NANOINFORMATICS: ADVANCING IN SILICO CANCER RESEARCH

by

David Eugene Jones

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

August 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **David Eugene Jones**

has been approved by the following supervisory committee members:

Julio Cesar Facelli	, Chair	**4/29/2016** Date Approved
Hamidreza S. Ghandehari	, Member	**5/2/2016** Date Approved
Bruce Earl Bray	, Member	**5/2/2016** Date Approved
John Franklin Hurdle	, Member	**4/29/2016** Date Approved
Karen Eilbeck	, Member	**5/2/2016** Date Approved

and by **Wendy W. Chapman**, Chair/Dean of the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Nanoinformatics is a relatively young field of study that is important due to its implications in the field of nanomedicine, specifically toward the development of nanoparticle drug delivery systems. As more structural, biochemical, and physiochemical data become available regarding nanoparticles, the greater the knowledge-gain from using nanoinformatics methods will become. While there are challenges that exist with nanoparticle data, including heterogeneity of data and complexity of the particles, nanoinformatics will be at the forefront of processing these data and aid in the design of nanoparticles for biomedical applications.

In this dissertation, a review of data mining and machine learning studies performed in the field of nanomedicine is presented. Next, the use of natural language processing methods to extract numeric values of biomedical property terms of poly(amido amine) (PAMAM) dendrimers from nanomedicine literature is demonstrated, along with successful extraction results. Following this is an implementation and its results of data mining techniques used for the development of predictive models of cytotoxicity of PAMAM dendrimers using their chemical and structural properties. Finally, a method and its results for using molecular dynamics simulations to test the ability of EDTA, as a gold standard, and generation 3.5 (G3.5) PAMAM dendrimers to chelate calcium.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF TERMS

Amber – Assisted Model Building with Energy Refinement, is both a package of molecular simulation programs and a set of molecular mechanical force fields intended for simulation of biomacromolecules.

Chelation – Refers to a type of bonding between ions/molecules and metal ions.

GAFF – General Amber Force Field.

Information Extraction – The computational task of extracting structured information from a corpus of unstructured and/or semi-structured documents.

J48 – A decision tree classifier, which is based on the C4.5 algorithm.

Machine Learning – A subfield of computer science whose origin evolved from the study of learning theory and pattern recognition in artificial intelligence.

Molecular Descriptors – The numerical representations of the structural properties of a given molecule, intended to be used mathematically to determine biochemical information regarding the molecule.

Molecular Dynamics (MD) Simulation – A computer simulation method for examining the theoretical physical movements of atoms and molecules.

Nanoinformatics – A field that was defined by the necessity to aid in the management and utilization of the vast amounts of data being produced by the field of nanomedicine, and more generally nanotechnology.

Nanomedicine – A field focused on the application of nanoscience techniques and nanoparticles to clinical research and healthcare, with the primary goal of using this technology for the prevention, diagnosis, and treatment of disease.

Nanoparticle Drug Delivery System – The use of a particular nanoparticle or nanoparticles as a vector for carrying and delivering a variety of payloads, including non-viral genes, small interfering ribonucleic acid, and cancer drugs, to combat viruses and cancer.

Natural Language Processing (NLP) – A field of computer science focused on the interaction of computers with human language.

Pharmacodynamics – The study of the effects that a drug has on the body or an organism.

Pharmacokinetics – The study of how the body affects the fate of a drug, examples include absorption, distribution, metabolism, excretion, and others.

Poly(amido amine) (PAMAM) Dendrimers – A class of nanoparticles distinguished by the highly branched structure, intended to be used as an oral therapy for cancer treatment and many other functions.

ACKNOWLEDGEMENTS

This work was supported by grant number T15LM007124 from the National Library of Medicine. Computational resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. Also, this work has been partially funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number 1ULTR001067, National Institutes of Health grants R01ES024681 and R01EB007470, and National Science Foundation grant CNS-1338155.

Special thanks to Julio C. Facelli for acting as head of my committee and being a great mentor, Hamid Ghandehari for participating as a member of my committee and providing his knowledge and expertise regarding nanomedicine, John F. Hurdle for participating as a member of my committee and providing his knowledge and expertise regarding natural language processing, and Bruce E. Bray and Karen Eilbeck for participating as members of my committee. Without all of your support, encouragement, and expertise this journey would have been much more daunting.

CHAPTER 1

INTRODUCTION

Nanomedicine

Nanoparticles, which are molecular particles with sizes ranging from 1 to 100 nanometers (nm), are being utilized in a variety of different fields. One field in particular is nanomedicine, which focuses on applying nanoscience techniques and nanoparticles to clinical research and healthcare, with the primary goal of using this technology for the prevention, diagnosis, and treatment of disease[1]. Diagnostic devices, tissue replacement, and pharmaceutics are just a few of the many applications for nanoparticles in the field of nanomedicine[2]. Nanoparticle drug delivery research currently being conducted in this field focuses on the use of nanoparticles carrying a variety of payloads, including nonviral genes, small interfering ribonucleic acids, imaging agents, tissue replacement therapies, and pharmaceutics, to combat viruses and cancer[3, 4]. This field is rapidly growing and evolving, as can be observed by the number of publications being produced on a yearly basis[1, 5]. This literature contains a vast amount of data with valuable knowledge regarding the relationship between the structure of nanoparticles (i.e., size, molecular weight, surface charge, and zeta potential) and their biological fate, which includes but is not limited to bioavailability and cytotoxicity[5].

Although nanoparticles possess the potential to be effective for treating disease, there are a number of challenges that exist. Several pharmacokinetic and pharmacodyamic (PK/PD) properties of nanoparticles, including absorption, distribution,

metabolism, excretion, and toxicity, are poorly understood and often differ substantially from traditional pharmaceutics[6]. PK/PD essentially focuses on how the body interacts with a drug, and the effect the drug has on the body. This dissertation focuses on two specific PK/PD properties, cytotoxicity and absorption of nanoparticles.

Cytotoxicity is an area of key concern to the nanomedicine community, because if a nanoparticle exhibits high cytotoxicity it is a definite cause for elimination from use in potential human applications[7, 8]. Several authors discuss the fact that toxicity of nanoparticles can be due to their cationic (positive) surface charge, which is necessary for interaction with the anionic (negative) cell membrane; if this charge is too great it can lead to cell membrane damage, degradation, and eventually cell lysis[7-10]. To counteract this cytotoxicity, synthetic methods can be utilized to engineer the surface of these nanoparticles with a variety of surface moieties, including biodegradable components. However, if the charge is completely neutralized or made anionic, it can result in a decrease in bioavailability[7-10].

When it comes to the pharmacokinetic property of absorption of nanoparticles, certain aspects are known and others are still unresolved. One of the major mechanisms for the oral absorption of nanoparticles is transcellular transport through the epithelium of the small intestines, specifically by way of endocytosis, and this transport can be influenced by many nanoparticle structural properties, such as surface charge and particle size[11]. Researchers have discovered a few mechanisms for enhancing the oral absorption of nanoparticles, intestinal wall bioadhesion, improved dissolution behavior, and transcellular uptake[12]. Due to the complexity and heterogeneity of the absorption mechanisms of nanoparticles, it is difficult to determine the predominant process by which transcytosis of nanoparticles occurs and identify specific attributes of these absorption mechanisms[13-15].

Due to the examples listed above, as well as many other potential challenges, the

development cycle of a nanoparticle drug delivery system for potential human applications can be difficult and lengthy. Tuning the many PK/PD properties to optimal configurations results in a nanoparticle synthesis and testing cycle that is quite repetitive and burdensome on scarce development resources. For these reasons, only a few nanoparticle systems have been used in approved FDA products to date[16, 17]. The nanomedicine community is very interested in the acceleration of translating nanomedicine bench scientific discoveries into clinical impact and practice. The ability to reliably predict and simulate PK/PD properties of nanoparticle drug delivery systems using *in silico* approaches has the potential for high payoff in nanomaterial development, allowing the concentration of scarce development resources into the synthesis and confirmatory testing of promising nanomaterials.

Nanoinformatics

To meet these ends and augment the field of nanomedicine, the field of nanoinformatics was established. Nanoinformatics is a field that was defined by the necessity to aid in the management and utilization of the vast amounts of data being produced in the field of nanomedicine, and more generally nanotechnology. The U.S. National Science Foundation laid the foundations for the field of nanoinformatics in 2007[18]. The 2014 National Nanotechnology Initiative Strategic Plan defines nanoinformatics as "the science and practice of developing and implementing effective mechanisms for the nanotechnology community to collect, validate, store, share, mine, analyze, model, and apply nanotechnology information"[19]. This same document also discusses the importance of establishing an improved nanoinformatics infrastructure, because it will improve the distribution and reproducibility of nanotechnology experimental data and promote the development and validation of models, tools, and techniques for transforming data into knowledge and applications[19]. Whereas

nanomedicine experimental research has well-developed protocols and methods, the corresponding nanoinformatics support for the field is much less developed and there is a substantial lack of authoritative sources of information accessible to noninformatics specialists[20]. Currently there are many major areas of informatics exploration being pursued in the field of nanomedicine, however this dissertation focuses on three: information extraction (IE), specifically natural language processing (NLP); data mining and machine learning; and molecular dynamics (MD) simulation. Several reviews have been published regarding research efforts in areas of nanoinformatics exploration, and for more information regarding the other areas of nanoinformatics research the reader should consult[1, 21-26].

The importance of using IE to rapidly advance biomedically relevant knowledge cannot be overstated, especially regarding domains where the literature corpus is diverse[27]. One particularly powerful IE technique is NLP, a group of methods focused on automatically extracting information from free-text and based upon semantic/syntactic analysis[28]. NLP methods have been very successful in a wide variety of biomedical domains, including nanomedicine[27, 29-31]. For more information regarding multiple implementations and uses of NLP in the field of nanomedicine, please refer to the review by Lewinski and McInnes[31].

Data mining and machine learning methods are quite often used to guide the design and development of small pharmaceutical compounds due to their success in medicinal chemistry[20]. One major focus of data mining and machine learning with this area of research is to establish quantitative structure activity relationships (QSARs), which link a molecule's structural properties to its functions. Increased use of predictive models within the field of nanomedicine is believed to lead to an acceleration in the translational process[1, 32, 33]. For more information regarding the application of data mining and machine learning to field of nanomedicine, please refer to Chapter 2, our review article regarding

the use of data mining and machine learning in the field of nanomedicine.

In small molecule drug delivery research, it is very common to see the use of MD simulations to gain understanding and test novel hypotheses at the molecular scale[34, 35]. Recent advances in computational power have made it so that these techniques can be applied in the field of nanomedicine. There are several publications regarding the use of MD simulations in the field of nanomedicine. Many of these publications examined binding and interactions between nanoparticles and several different biological molecules[36-38]. Other publications have examined endocytosis of a variety of nanomedicines and biocorona formation on silver nanoparticles[39, 40].

PAMAM Dendrimers

Poly(amido amine) (PAMAM) dendrimers were selected as the nanoparticle drug delivery system of interest for this dissertation because they are well documented, have the potential to be used as successful delivery vectors, and can be delivered orally rather than intravenously[41]. These particular nanoparticle drug delivery systems are well-defined, highly branched structures consisting of a central core, typically ethylene diamine, surrounded by concentric shells, amido amine branches[42, 43]. As a naming convention, the number of concentric shells that surround the core of a PAMAM dendrimer is used to determine the generation of that PAMAM dendrimer. Amine- and/or hydroxyl-terminated PAMAM dendrimers are considered to be the full generation dendrimers and named as generation 1, 2, 3, etc. Carboxyl-terminated PAMAM dendrimers are considered to be the half generation dendrimers and named as generation 1.5, 2.5, 3.5, etc. The structure of these polymeric nanoparticles can be modified easily to a variety of specifications, and due to their scaffold-like nature, they are capable of carrying a variety of bioactive agents and improving the solubility of poorly soluble ones[44, 45]. These properties make PAMAM dendrimers promising candidates as drug carriers.

Despite the promising attributes of PAMAM dendrimers, a significant barrier for their use in human applications exists due to their potential toxicological effects, depending upon the structure of PAMAM dendrimer that is used. Research has shown that cationic PAMAM dendrimers can exhibit generation, concentration, and surface-charge-dependent toxicity[46-49].

Very little research has been conducted in applying nanoinformatics methods and techniques toward PAMAM dendrimers. Other than the chapters presented in this dissertation, much of the nanoinformatics focus regarding PAMAM dendrimers has been MD simulations. One data mining paper examined the ability to predict embryonic zebrafish postfertilization toxic effects of several nanoparticles, including metal nanoparticles, dendrimers, metal oxides, and polymeric materials[50]. The MD simulation studies regarding PAMAM dendrimers have mostly focused on gaining insight on PAMAM dendrimer/ligand conformations and energies. The molecular docking studies have analyzed interactions between PAMAM dendrimers and several molecules (siRNA, curcumin, porphyrin, and pharmaceutical agents)[51-56]. Other PAMAM dendrimer MD simulation research has examined bivalent binding, design of multi-functional PAMAM dendrimer-based nano-therapeutics, and identification of key structural design principles for bioactive dendrimer molecules[57-59].

Motivation and Objectives

The ability to reliably predict and simulate PK/PD properties of orally delivered PAMAM dendrimer nanoparticle drug delivery systems using *in silico* approaches has the potential for high payoff in nanomaterial development, allowing the concentration of scarce development resources into the synthesis and confirmatory testing of promising PAMAM dendrimers. Due to the challenges and knowledge gaps that exist with applying PAMAM dendrimers toward human applications, this figured to be a valuable

nanoparticle model to apply nanoinformatics methods and techniques. Our belief is that the applications of nanoinformatics towards PAMAM dendrimers will function as a proof-of-concept model that can be expanded to a variety of nanomedicines.

This dissertation contains research demonstrating the use of nanoinformatics methods and techniques, specifically NLP, data mining and machine learning, and MD simulation, to assess PAMAM dendrimers. Background material, four articles written for submission to academic journals, and a discussion of this research make up this dissertation. The coming chapters present the work that was undertaken in an effort to reliably predict and simulate the PK/PD properties of cytotoxicity and calcium chelation of orally delivered PAMAM dendrimers. Chapter 2 is a review article of data mining and machine learning studies performed in the field of nanomedicine. This chapter examines the variety of nanoparticle properties that have been predicted using data mining and machine learning within the field of nanomedicine. Chapter 3 is a journal article discussing the use of natural language processing methods to extract numeric values of biomedical property terms of poly(amido amine) (PAMAM) dendrimers from nanomedicine literature. The goal is to successfully extract this numeric data to be utilized in the subsequent journal articles. Chapter 4 is a journal article describing the development of a predictive model of cytotoxicity of PAMAM dendrimers on human colon carcinoma (Caco-2) cells using their chemical and structural properties. This establishes a QSAR for developing nontoxic PAMAM dendrimers. Chapter 5 is a journal article analyzing the use of MD simulations to test the ability of generation 3.5 (G3.5) PAMAM dendrimers to chelate calcium as a potential mechanism for absorption in the intestine. Chapter 6 contains a discussion of the importance of research findings from each journal article, the contribution of this work to the field of biomedical informatics, and potential for future research.

CHAPTER 2

A REVIEW OF THE APPLICATIONS OF DATA MINING AND MACHINE LEARNING FOR THE PREDICTION OF BIOMEDICAL PROPERTIES OF NANOPARTICLES[1]

Abstract

The field of nanomedicine is increasingly becoming a very active field of research. However, information about exciting nano-QSAR approaches is not readily available to noninformatics specialists interested in nanomedicine. The goals of this review are to: (1) review research involving the use of data mining and machine learning for the prediction of biomedical properties of nanoparticles of medical interest and (2) examine the progress and challenges that this relatively new field of nanoinformatics faces to become a major contributor toward the development of effective nanomedicines. A comprehensive search of the existing literature in the field of nanomedicine referencing the use of data mining and/or machine learning techniques was conducted in the fall of 2015. The search produced papers that include a large and varied number of data mining applications to predict biomedical properties of nanomaterials of medical interest.

This article presents a comprehensive review of applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles of medical interest. These include the influence of particle physiochemical properties on cellular

[1] *Reprinted from Computer Methods and Programs in Biomedicine, Copyright 2016. David E. Jones [a], Hamidreza Ghandehari [b, c], and Julio C. Facelli [a, c]. [a] Department of Biomedical Informatics, [b] Departments of Bioengineering and Pharmaceutics and Pharmaceutical Chemistry, and [c] Utah Center for Nanomedicine, Nano Institute of Utah, University of Utah, Salt Lake City, UT 84112, US*

uptake, cytotoxicity, molecular loading, and molecular release, in addition to manufacturing properties like nanoparticle size and polydispersity, which can be predicted using data mining and machine learning methods. Overall the results are encouraging and suggest that as more systematic data from nanoparticles become available, machine learning and data mining would become a powerful aid in the design of nanoparticles for biomedical applications. There is, however, the challenge of great heterogeneity in nanoparticles, which will make these discoveries more challenging than for traditional small-molecule drug design.

Introduction

The field of nanomedicine, which focuses on the use of nanoparticles and nanotechnology in the bio-medical domain, is increasingly becoming a very active field of research. To date only a few nanoparticle systems have been used in FDA-approved products[16, 17], and there is a great deal of interest in accelerating the translation of nanoscience bench scientific discoveries into clinical practice. While nanomedicine research has well-developed experimental protocols, the corresponding informatics support for nanomedicine is less developed and there is a substantial lack of authoritative sources of information accessible to noninformatics specialists[20]. Increasing the use of nanoparticle quantitative structure activity relationships (nano-QSARs)[32, 60] and other predictive models in the field of nanomedicine can greatly accelerate the translational process[1, 32, 33]. However, information about exciting nano-QSAR approaches is not readily available to noninformatics specialists interested in nanomedicine. The goals of this review are to: (1) review research involving the use of data mining and machine learning for the prediction of biomedical properties of nanoparticles of medical interest and (2) examine the progress and challenges that this relatively new field of nanoinformatics faces to become a major contributor toward the development of effective nanomedicines.

A comprehensive search of the existing literature in the field of nanomedicine referencing the use of data mining and/or machine learning techniques was conducted in the fall of 2015. Both Scopus and PubMed were accessed using the search criteria, "nanomedicine AND ((data mining) OR (machine learning))." Upon retrieving the initial set of articles, they were reviewed to assess content as well as to gather additional references from related publications. The methods and results reported in these publications are discussed in the following sections, while in the discussion section we present the authors' perspective about the successes and remaining challenges when using artificial intelligence and data mining for the prediction of biomedical properties of nanoparticles of medical interest.

The research papers covered in this review focus on applications of data mining and machine learning to nanoinformatics, with the goal of developing predictive models for a variety of nanoparticle properties and their biological effects. The material is divided into two main sections, one discussing papers in which some of the biomedical effects of nanoparticles are predicted using nanoparticle properties/conditions, and the other one discussing papers that, using the aforementioned techniques, attempt to predict actual molecular or aggregate properties of nanomaterials based on their composition and processing.

Predicted Properties

In this section the material is organized into sections covering the properties discussed above as follows: cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanoparticle size, and polydispersity.

Cellular Uptake

Significant efforts have been made in the field of nanomedicine to understand and improve cellular uptake and targeting. This is primarily driven by the desire to use

nanoparticles to treat cancer by using them to deliver biologically active compounds specifically to cancerous cells[61-64]. The use of predictive models and nano-QSARs in this area could be very beneficial because the development cost of novel nanoparticles with the desired properties is quite high and the design space is quite large. Any computational tool that can assist in reducing the design space by quantitatively predicting the characteristics of desirable molecules before synthesis, would allow researchers to dedicate limited resources toward performing experimental work on the most promising candidates.

Two papers have reported the use of data mining and machine learning to predict cellular uptake of nanoparticles. Both papers examined the cellular uptake data of cross-linked iron oxide (CLIO) nanoparticles from the paper by Weissleder[63].

Fourches *et al.*[33] developed a method for predicting the cellular uptake of CLIO nanoparticles, with a variety of small organic molecules decorating their surface, by human pancreatic cancer cells (PaCa2) as a function of the nanoparticle properties. For the 109 organic compounds in their study, they calculated 150 two-dimensional MOE descriptors (using commercial software distributed by the Chemical Computing Group), which included surface areas, physical properties, Kier & Hall connectivity indices, kappa shape indices, atom and bond counts, adjacency and distance matrix descriptors, molecular charges, and pharmacophore feature descriptors. Their method utilized a 5-fold cross validation *k* Nearest Neighbors (*k*NN) regression as the prediction algorithm. *k*NN is an algorithm whose central concept is that the activity of a certain compound can be predicted by examining the average activities of *k* compounds from the dataset that share chemical similarity with the compound[65, 66] under consideration. Initial results for their model showed an R^2 value of 0.72, but when they applied an applicability domain criterion and removed compounds that were outside of the domain under consideration, the R^2 value improved to 0.77. They found that the most important features in this model are associated with the lipophilicity of the compounds. In both models, the observation was

made that the more lipophilic the molecule bound to the CLIO nanoparticle, the greater the cellular uptake.

Winkler *et al.*[67] examined the ability of machine learning techniques to predict cellular uptake of CLIO nanoparticles, decorated also with a variety of small molecules, by human umbilical vein endothelial cells (HUVEC) and PaCa2. The dataset consisted of 108 samples, which were split into 21 samples for the test set and 87 samples for the training set. Two-dimensional DRAGON descriptors were calculated for the decorated CLIO nanoparticles[68] and two different models were developed, a linear and nonlinear nano-QSAR model. A multiple linear regression algorithm along with an expectation minimization algorithm with a sparse (Laplacian) prior were used to develop the sparse linear nano-QSAR model[69]. A Bayesian regularized neural network with either a Gaussian or Laplacian prior was used to develop the sparse nonlinear nano-QSAR model[70-72]. The linear and nonlinear models for HUVEC uptake utilized 11 of the DRAGON descriptors and yielded R^2 values of 0.63 and 0.66, respectively. The linear and nonlinear models for PaCa2 uptake utilized 19 of the DRAGON descriptors and yielded R^2 values of 0.79 and 0.54, respectively.

The authors reported little to no overlap in the sets of DRAGON descriptors used for the HUVEC and PaCa2 cellular uptake models, and concluded that this suggests that different mechanisms for cellular uptake may be utilized by these two cell types[73]. The authors of this paper suggest that poor results observed in the macrophage and macrophage-like cells may be due either to the small size or surface modifications of the nanoparticles used in this study.

Both of these studies, Fourches *et al.*[33] and Winkler *et al.*[67], reported similar results for their best performing method to predict PaCa2 cell line uptake (0.77 vs. 0.79), even though they used slightly different predictive methods and molecular descriptors. More than likely the reason for this is that they used the same dataset of 109 fluorescent

nanoparticles taken from a study performed by Weissleder[63].

Cytotoxicity

Predicting cytotoxicity of nanoparticles has been the most common application of data mining and machine learning to research in nanoinformatics. Cytotoxicity to non-target cells is a major concern in nanomedicine[7, 8] because the use of nanoparticles for human treatment is contingent upon low cytotoxicity of the carriers at the needed therapeutic doses. Toxicity is also a serious concern for nanoparticles used in consumer products due to their potential environmental impact[22]. The ability to predict cytotoxicity via *in silico* approaches is highly desirable because of the potentially high payoff in nanomaterial design and the development of prescreening for toxicity. This can result in shifting limited development resources into the synthesis and testing of nanoparticles with predicted low cytotoxicity[74] that are more likely to be suitable for human treatment or consumption.

Experimentally, cytotoxicity can be measured by a number of *in vitro* toxicity assays that can infer cytotoxicity by examining different cellular parameters, including but not limited to oxidative stress, inflammatory response, genotoxicity, and cell viability[75]. The articles reviewed in this area of data mining and machine learning research report results on several different nanoparticle types, cell types, and cytotoxicity analysis methods. A summary of the systems studied, methods, and findings from these research articles is given in Table 2.1.

Many of the articles reported in Table 2.1 report cytotoxicity prediction of metal oxide nanoparticles. Sayes and Ivanov[76] used linear discriminant analyses (LDA) classification[77] and multivariate linear regression to predict lactate dehydrogenase (LDH) release from rat lung alveolar macrophages and immortalized rat L2 lung epithelial cells, caused by exposure to titanium dioxide (TiO_2) and zinc oxide (ZnO) nanoparticles[76]. LDH

release is indicative of cell membrane damage and ultimately cell death. The TiO_2 nanoparticles were characterized by five different physiochemical properties that were experimentally measured and then used as feature descriptors in the models, these properties are the following: engineered size, size in water, size in PBS, concentration, and zeta potential. The ZnO nanoparticles were characterized by six different physiochemical properties that were experimentally measured: engineered size, size in water, size in PBS, size in CCM, concentration, and zeta potential. For both the TiO_2 and ZnO nanoparticles, all possible combinations of descriptors were analyzed for the predictive models. The dataset consisted of a total of 42 samples, 24 TiO_2 nanoparticle samples at different concentrations ranging from 25-200 mg/L and 18 ZnO nanoparticle samples at different concentrations ranging from 25-100 mg/L. TiO_2 and ZnO nanoparticle sample sets were analyzed independently. The multivariate linear regression analysis of the TiO_2 nanoparticles yielded R^2 values ranging from 0.15-0.70, with the highest performance being observed when all possible descriptors were utilized, which may be an indication of overfitting. The LDA analysis of the TiO_2 nanoparticles yielded R^2 values ranging from 0.70-0.77. Due to the observed correlation between the different size measurements for ZnO, only all possible combinations of engineered size, concentration, and zeta potential were examined in the multivariate linear regression analysis predictive models of ZnO. The analysis yielded R^2 values ranging from 0.19-0.49, with the highest performance model being obtained when all these descriptors were utilized, leading the authors to conclude that their dataset did not have enough data to obtain accurate predictions of LDH release for ZnO. The authors also acknowledge that for ZnO there might be other features that were not present in their dataset that are necessary to obtain better prediction models. This highlights the need of larger well curated datasets to gain a better understanding of the real limitation of nano-QSAR methods.

Puzyn *et al.*[60] predicted the cytotoxicity, specifically the effective concentration of

a compound that causes bacterial viability to be reduced by 50%, EC_{50}, of *Escherichia coli* (*E. coli*), caused by exposure to 17 different metal oxide nanoparticles: ZnO, CuO, V_2O_3, Y_2O_3, Bi_2O_3, In_2O_3, Sb_2O_3, Al_2O_3, Fe_2O_3, SiO_2, ZrO_2, SnO_2, TiO_2, CoO, NiO, Cr_2O_3, and La_2O_3. The MOPAC 2009 software package was used to calculate 12 different molecular descriptors (standard heat of formation of the oxide cluster, total energy of the oxide cluster, electronic energy of the oxide cluster, core-core repulsion energy of the oxide cluster, area of the oxide cluster calculated, volume of the oxide cluster calculated, energy of the highest occupier molecular orbital of the oxide cluster, energy of the lowest unoccupied molecular orbital of the oxide cluster, energy difference between HOMO and LUMO energies, enthalpy of detachment of metal cations Me^{n+} from the cluster surface, enthalpy of formation of a gaseous cation, and lattice energy of the oxide) of the metal oxide nanoparticles. A multiple regression method was combined with a genetic algorithm to find the best model for the prediction of cytotoxicity[78]. Selection of the best combination of calculated descriptors was performed by the genetic algorithm, which found that the enthalpy of formation of a gaseous cation having the same oxidation state as in the metal oxide structure, ΔH_{Me+}, is the best descriptor. The multiple regression using this descriptor reached an R^2 value of 0.85, with an externally validated regression coefficient, Q^2_{ext}, of 0.83, and an RMS error of 0.19. The authors concluded that their model can be used to predict the toxicity of novel, untested metal oxide nanoparticles, but this only applies if the structure is not significantly different from the metal oxide nanoparticles in the training set, limiting the generalizability of this approach.

Using logistic regression models, Liu *et al.*[79] classified cytotoxicity by examining the plasma membrane integrity when transformed bronchial epithelial cells (BEAS-2B) were exposed to nine different metal oxide nanoparticles (Al_2O_3, CeO_2, Co_3O_4, TiO_2, ZnO, CuO, SiO_2, Fe_3O_4, and WO_3). For the development of the model, a set of 10 nanoparticle descriptors was selected and measured experimentally. These descriptors include simple

constitutional descriptors (number of oxygen atoms in the metal oxide, number of metal atoms in the metal oxide, metal oxide molecular weight, and atomic mass of the metal); stability and reactivity information (atomization energy); element group properties (periodic table group and period of the metal in the metal oxide); simple geometric descriptor (nanoparticle primary size); and indicators of surface charge and aggregation tendency (zeta potential and isoelectric point). Additional experimental conditions were taken into account by adding measured values for a set of four different concentrations as input parameters of the model. The paper does not specifically state the number of samples used in the dataset, however, it appears that 83 samples were used. All possible combinations of the descriptors and concentrations were analyzed for their nano-QSAR models, which generated accuracies ranging from 93 to 100%. The atomization energy of the metal oxide, nanoparticle size, nanoparticle volume fraction, and period of the metal in the nanoparticle were the four descriptors used in the best performing model. The authors observed that the atomization energy had the greatest contribution to the model, showing that as the atomization energy decreases, the toxicity of the metal oxide nanoparticle increases. They argue that this could be explained by the decrease in stability of the metal oxide nanoparticle and the increase of its reactivity. The authors of this paper were impressed by their results but stated that it is necessary to expand the experimental dataset used in order to increase confidence and improve the reliability of their results, as the high accuracies reported may be a consequence of either overfitting or perhaps lack of diversity in the reference data.

Horev-Azaria *et al.*[80] used a J48 classification model to predict cytotoxicity, measured as cell viability using a binary classification of toxic or nontoxic, of cobalt ferrite nanoparticles on seven different cell lines (A549, NCI H441, HepG2, MDCK, Caco-2 TC7, TK6, and primary mouse dendritic-cells) and precision-cut rat lung slices. J48 is a decision tree classifier, which is based on the C4.5 algorithm[81]. The paper does not specifically state

the number of samples used in the dataset, but it appears that 151 samples were used. Their model involved the use of a ten-fold cross-validation that was tested for three different decreasing cell viability values: 30%, 25%, and 20%. The accuracy of their model reached 92.5%, 89%, and 85.2% for the 30%, 25%, and 20% cell viability values, respectively. The J48 decision tree shows that the most important descriptor used in making the predictions was the concentration of cobalt ferrite nanoparticles. Also two experimental conditions were present in the decision tree for making the predictions, cell-type and exposure time, but no intrinsic nanoparticle properties were found of importance in the model. The authors of this paper indicated that their study is restricted to a specific type of nanoparticle and the cell lines used, and to make this model more generalizable, it would require a significantly larger database of different nanoparticles and cell lines.

Winkler *et al.*[67] reported the use of Bayesian neural networks and multiple linear regression models to predict smooth muscle cell apoptosis caused by 50 different CLIO nanoparticles on endothelial muscle cells, smooth muscle cells, hepatocytes, and monocytes. The data also consist of four biological assays to determine toxicity and four different concentrations of nanoparticles used, yielding a sample size of 3,200 samples. The two models utilized here are presented above in the cellular uptake section of the review. The linear and nonlinear models tested in this work achieved R^2 values of 0.86 and 0.90, respectively. The authors observed that the nanoparticles' core material, surface coating type, and surface charge were the most important features needed to make accurate predictions of the smooth muscle apoptosis caused by CLIO nanoparticles.

Using a Monte Carlo method, Toropova *et al.*[82] built a nano-QSAR model to predict pLC50 values, which are the negative decimal logarithm of the lethal concentration of nanoparticle that causes 50% of the original bacterial population to die, induced by metal oxide nanoparticles in *E. coli*. The paper does not specifically state the number of samples used in the dataset. The authors utilized quasi-SMILES as their calculated descriptors for

the model. The data for their study was split into six different datasets, where each dataset was used as a training, calibration, or testing set. Their models yielded R^2 values ranging from 0.73 to 0.98. The authors of this paper state that their method of distributing data into training, calibration, and validation sets significantly influences the results of their study, and suggest that data should be distributed into training and external validation sets instead to improve reliability of the study.

Fourches *et al.*[33] used support vector machine-based classification to predict cytotoxicity as a binary value (toxic/nontoxic) of a variety of metal nanoparticles (CLIO, pseudocaged, monocrystalline iron oxide, CdSe core quantum dots, and iron-based) on four different cell lines (monocytes, hepatocytes, endothelial cells, and smooth muscle cells). Their model utilized experimentally measured attributes, describing the nanoparticles (nanoparticle size, zeta potential, and relaxivity, which represents the magnetic properties of the nanoparticle). The biological activity profile was represented by the dose, cell line, and assay utilized for the nanoparticle; all of these values were used to create an arithmetic mean which was then used to create the binary classification to be used by the model as toxic or nontoxic. The number of samples used in this study is not explicitly stated, but it can be estimated to be 3,264 samples. A five-fold cross-validation method was used for their model, and prediction accuracies ranging from 56% to 88% were achieved by their model. The authors of this paper suggest that exploration of many different approaches will be necessary to identify and predict relationships between metal nanoparticle structures and their biological activity in order to provide more generalizable nano-QSAR relationships.

Liu *et al.*[50] used a variety of algorithms (IBK, Bagging, M5P, and KStar) in an effort to predict embryonic zebrafish postfertilization toxic effects of several nanoparticles, including metal nanoparticles, dendrimers, metal oxides, and polymeric materials. IBK is a K-nearest neighbor predictor that assigns an input to the most common output label

among its K nearest neighbors[83]. Bagging is a hybrid classification method that creates classes and reduces variance by bagging classifiers[84]. M5P is a tree algorithm that generates M5 Model trees and rules[85], and KStar is an instance-based classifier where the test instance's class is based upon the class of similar training instances[86]. The paper does not specifically state the number of samples used in the dataset. Their model used 20 input variables representing the nanoparticle properties (e.g., particle size distribution, structure, surface charge, water solubility, etc.) and experimental conditions (e.g., exposure route, concentration, duration, etc.). The most successful predictions were obtained for the 24-hour postfertilization mortality, 120-hour postfertilization mortality, and 120-hour postfertilization heart malformation, with accuracies of 0.84, 0.77, and 0.73, respectively, when using the IBK algorithm. For other prediction models, the accuracies corresponding to the prediction of different properties using a variety of methods are reported in Table 2.2. The results of Liu *et al.*[50] indicate that dosage concentration, shell composition, and surface charge are the most important attributes when analyzing embryonic zebrafish postfertilization mortality, which agrees with previous bench studies[87, 88]. The authors of this paper discussed increasing the size and diversity of the data used for their study to expand and refine the impact of their predictive models.

Jones *et al.*[74] tested the ability of a variety of algorithms (Naïve Bayes, SMO, J48, Bagging, Classification via Regression, Filtered Classifier, LWL, Decision Table, DTNB, NBTree, and Random Forest) to predict the cytotoxicity, measured as cell viability considered as a binary variable (toxic/non-toxic) of poly(amido amine) (PAMAM) dendrimers on human colorectal cancer cells (Caco-2). Naïve Bayes is a Bayesian classifier which uses posterior probability to predict the value of the target attribute[89], e.g., by using a given input attribute, the classifier attempts to find the target attribute value that maximizes the conditional probability of the target attribute. SMO is a support vector machine classifier that globally replaces all values and transforms nominal attributes into

binary ones[90]. This method starts with large sets of cases which belong to known classes, cases are analyzed for patterns that allow for reliable discrimination of classes. The patterns are represented as models, either in the form of decision trees or sets of if-then rules which can be used to classify new cases. Classification via regression performs its classification by binarizing each class and building one regression model for each class[91]. Filtered classifier is an arbitrary classifier that runs on data passed through an arbitrary filter[92]. LWL uses an instance-based algorithm to assign instance weights, the abbreviation stands for locally weighted learning[93]. Decision table is a simple decision table majority classifier[94]. DTNB is a decision table/Naïve Bayes hybrid classifier. During the search the algorithm determines the need to divide the attributes into two disjoint subsets: one for the decision table, the other for Naïve Bayes[95]. NBTree is a decision tree/Naïve Bayes hybrid classifier that builds a decision tree with Naïve Bayes classifiers at the leaves[96]. The dataset used in this study consisted of 103 samples. The predictive models utilized 51 molecular descriptors (e.g., molecular weight, pI, molecular polarizability, etc.), which were calculated using MarvinSketch[97]. Their models achieved 10-fold cross-validated accuracies ranging from 65.0 to 83.5%. Their best classification models were obtained using the J48 and Bagging methods with 10-fold cross-validated accuracies of 83.5% for both methods. The decision tree from their J48 classifier (see Figure 2.1) shows that the descriptors used in making the best prediction were pI, molecular weight, and concentration of PAMAM dendrimer. Indications of the importance of using larger datasets to create more reliable and robust classification models are made by the authors of this paper.

Overall, there is a tendency toward developing models for toxicity of metal oxide and metal nanoparticles. There are still many classes of nanoparticles for which no work has been reported on the use of data mining methods to predict their cytotoxicity, such as micelles, liposomes, polymeric nanoparticles, etc. It is encouraging to see that many of the

papers concluded that properties related to charge, concentration, and size of nanoparticles are important in developing predictions of cytotoxicity. These properties have been hypothesized to be important indicators of the potential cytotoxicity of nanoparticles[98], but the results compiled in this review provide substantial computational verification. The collection/aggregation of more data regarding cytotoxicity is a definite must for the further development of cytotoxicity prediction methods of nanoparticles.

Molecular Loading

Molecular loading is a very important property for nanoparticles when these nanoparticles are intended to be used as delivery devices of drugs and/or image contrasting agents to specific tissues or cells. Two research articles reported the use of data mining and machine learning techniques to predict the ability to load molecules into nanoparticles. A summary of the findings from these research articles is given in Table 2.3.

Winkler *et al.*[67] explored the use of Bayesian neural networks and multiple linear regression models to predict inhibition of acetylcholinesterase (AChE) or nonspecific adsorption, and nonspecific protein binding to surface-modified gold nanoparticles. The dataset used for this study consisted of 80 samples. Two-dimensional DRAGON descriptors were calculated for the surface-modified gold nanoparticles [68]. The two models utilized here are presented above in the Cellular Uptake section of the review. The linear and nonlinear models for AChE inhibition utilized 14 of the DRAGON descriptors and yielded R^2 values of 0.81 and 0.80, respectively. The linear and nonlinear models for nonspecific protein binding utilized 10 of the DRAGON descriptors and yielded R^2 values of 0.93 and 0.94, respectively. The nonspecific protein binding correlated only with the concentration of protein and did not exhibit any dependence on the nanoparticle properties. The authors state that for their model the results were good, but care should be taken when making new predictions because there is a need for reasonably sized

datasets, high-quality data, and high-quality descriptors to further verify the generalizability of their models.

Shalaby *et al.*[99] used artificial neural networks (ANN) to predict the entrapment efficiency of noscapine in di- and triblock co-polymers of poly(ethylene glycol) and poly(lactide). The number of samples used in this study is not explicitly stated. The experimentally measured input variables used by their model were the molecular weight of the polymer, the ratio of polymer to drug, and number of blocks per polymer. Their model yielded an overall R^2 value 0.91486 for entrapment efficiency of noscapine predictions. The ratio of polymer to drug was the most important feature for the predictions of noscapine entrapment efficiency. Experimentally, similar results have been seen in poly(lactide) (PLA) and poly(ethylene glycol) (PEG)[100].

Both research articles showed promising results in the use of data mining and machine learning to predict the ability to load molecules to nanoparticles. The two articles reported R^2 values above 0.90, and showed clear evidence of first-order reaction dynamics for the entrapment process that as the most important feature for predicting the ability to load molecules to nanoparticles correlates with the amount of the molecules available in solution to be loaded. Clearly more data regarding a much more diverse number of nanoparticles are needed to evaluate the relevance of nanoparticle property on molecule loading beyond their concentration.

Molecular Release

Due to the toxic nature of many cancer drugs, it is important to encapsulate or conceal them within a nanoparticle until they reach the cancerous cells or target tissues in the body[101, 102]. Some nanoparticle drug delivery systems also possess the ability to be suitable carriers for unstable active pharmaceutical ingredients[103] in order to protect them from degradation before reaching the release target cells or tissues. Two publications

examined the use of data mining and machine learning to predict the ability to release molecules encapsulated in nanoparticles. A summary of the findings from these research articles is given in Table 2.4.

Husseini *et al.*[101] used ANN to model the release of doxorubicin from polymeric (Pluronic P105) micelles at two different frequencies of ultrasound. The number of samples used in this study is not explicitly stated. The model was trained using experimentally obtained input-output data of the release of doxorubicin from the micelles. The predictions made by the ANN method corresponded closely to the experimental data used, and the maximum prediction errors at the ultrasound frequencies of 20 and 70 kHZ were 0.002 and 0.001, respectively.

Szlek *et al.*[104] used ANN to predict the release of macromolecules (bovine serum albumin, human serum albumin, recombinant human erythropoietin, lysozyme, recombinant human epidermal growth factor, recombinant human growth hormone, beta-amyloid, recombinant human erythropoietin coupled with human serum albumin, hen ovalbumin, insulin, bovine insulin, L-asparaginase, chymotrypsin, and alpha-1 antitrypsin) from poly(lactic-co-glycolic acid) (PLGA) nanoparticles. The paper does not specifically state the number of samples used in the dataset, but it appears that 754 samples were used with 320 variables. The independent parameters used in the models included formulation characteristics, experimental conditions, and molecular descriptors calculated using the Marvin cxcalc plugin[105]. Feature selection was performed in order to remove features that did not improve the predictions, and resulted in four different analyses. These models achieved relative root-mean-squared errors of 17.7, 17.1, 16.4, and 15.4 when using 21, 17, 16, and 11 features as input variables, respectively, and using the monotone multilayer perceptron neural network. The analysis with eleven feature input variables was the best and included Szeged index, pI, quaternary structure of macromolecule, lactide-to-glycolide in polymer ratio, poly(vinyl alcohol) inner phase

concentration, poly(vinyl alcohol) outer phase concentration, encapsulation rate, mean particle size, dissolution pH, production method, and percentage of macromolecule dissolved as the input variables.

As can be seen from the results, these two articles demonstrate that it is feasible to create predictive models for the quantitative release of molecules from nanoparticles. Several different molecules released and nanoparticles were studied, however it is necessary to evaluate a wider variety of nanocarriers, predictive algorithms, and carried substances to make a final determination of the power of machine learning for this application.

Nanoparticle Adherence

Often in the treatment and imaging of cancer, and to take advantage of the enhanced permeability and retention of smaller nanoparticles, researchers limit the size of synthesized nanoparticles to 200-300 nm[106]. This is not always necessarily the best strategy for development of new therapies because there are many limitations that exist for enhanced permeation and retention-based therapies[107]. The nanoparticles in the study discussed below were designed to adhere to the walls of diseased blood vessels and avoid dislodgement from hydrodynamic forces and provide a useful data set to explore data mining and machine learning to predict nanoparticle adherence.

Boso *et al.*[108] utilized ANN to predict the number of polystyrene fluorescent nanoparticles adhering to the vessel walls as a function of wall shear rate and nanoparticle diameter. This is important because it is desired to develop an optimal structural configuration of nanoparticles to enhance their accumulation in diseased tissues. The paper does not specifically state the number of samples used in the dataset. The ANN performed quite well at predicting the optimal particle diameter with root mean squared error values of 0.03678 nm and 0.03460 nm, respectively. The authors of this article claim

that this work demonstrated that by using ANN, the number of long parallel plate flow chamber experiments can be minimized due to the accuracy of the predictive models. They also argue that the predictive model developed could be optimized for *in vivo* studies, thereby limiting the amount of animal experimentation.

Nanoparticle Size

As can be seen above, the size of nanoparticles is a very important property that can affect their usefulness in nanomedicine, for instance the size of a nanoparticle has been found to be a very important factor determining the fate of the nanoparticle *in vivo*[109]. Optimization of size is also important for the design and development of nanoparticles used to treat a variety of tumors, because the size of the nanoparticle affects their permeability and retention[106]. Nanoparticle size can change based upon solution conditions, manufacturing, drug loading, and release of drugs[110]. Two publications examined the use of data mining and machine learning to predict nanoparticle size. A summary of the findings from these research articles is given in Table 2.5.

Asadi *et al.*[111] determined the features that are most relevant to the prediction of particle size of tri-block poly(lactide)-poly(ethylene glycol)-poly(lactide) (PLA-PEG-PLA) nanoparticles using ANN. The paper does not specifically state the number of samples used in the dataset, but it appears that 51 samples were used. There were four input variables used in this study: amount of drug, nanoparticle polymer concentration, mixing rate, and solvent ratio. The method predicted the size of the polymer-based nanoparticles in nm. The specific ANN used was a three-layered feed-forward back propagation neural network, and it achieved an R^2 value of 0.9434 for the validation data. They found that the nanoparticle polymer concentration is the most important feature when determining nanoparticle size.

Shalaby *et al.*[99] used ANN to predict the particle size of di- and triblock copolymers

of poly(ethylene glycol) and poly(lactide). The paper does not specifically state the number of samples used in the dataset, but it appears that 27 samples were used. There were three input variables used in this study: nanoparticle polymer molecular weight, ratio of nanoparticle polymer to drug, and number of blocks in the nanoparticle copolymer. The method predicted the size of the polymer-based nanoparticles in nm. Their model yielded an R^2 value of 0.9783. The prediction of particle size was mostly determined by the molecular weight of the nanoparticle polymer type.

Clearly, the methods for predicting nanoparticle size are quite accurate. One potential reason for this observation is that for these methods the prediction of the nanoparticle size appears to be dependent upon only a single feature. However, this single feature is different for the different cases reported in the articles listed above.

Polydispersity

One of the many challenges and goals of the field of nanomedicine is the ability to prepare narrowly dispersed nanoparticles[112]. Commonly, nanoparticles exhibit relatively high polydispersity that can result in a number of drawbacks, like mixture of nanoparticles with varying loading capacities, decrease in physical stability, variety of release profiles, and unpredictable degradation and clearance rates[113-116].

Esmaeilzadeh-Gharehdaghi *et al.*[117] predicted the polydispersity of chitosan nanoparticles using ANN with four input features: amplitude of sonication of chitosan solution, sonication time of chitosan solution, chitosan solution concentration, and chitosan solution pH. The dataset used in this study consisted of 39 samples. The application of the model to the validation data yielded an R^2 value of 0.84. The data mining work revealed that when the chitosan solution concentration was increased, polydispersity decreased, and that when the pH of the chitosan solution is lower or more acidic, the polydispersity increases.

Discussion

The steady growth of the field of nanomedicine has led to the development of nanoinformatics and subsequently the use of data mining and machine learning to develop nano-QSARs and other methods to predict both functional and structural properties of nanoparticles. Research articles focusing on this area of research appear to be published in a wide variety of journals. The methods reported attempt to predict a large number of nanoparticle properties, including cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanoparticle size, and polydispersity.

There are two common themes that can be observed from the papers reviewed here. First, the most common method used to create predictions is some variant of artificial neural networks, ANN. There are several reasons for which this may be considered the method of choice, including the complexity of nanoparticle data, large number of attributes describing nanoparticles, and the potential difficulty in creating a prediction with rule-based algorithms due to the lack of sufficient empirical knowledge. The most common descriptors or attributes necessary to create accurate predictions often involve charge, concentration, and size-based properties of nanoparticles.

Cytotoxicity from inorganic materials is the most commonly predicted nanoparticle property, and most reports find that the most common factors determining it are charge, concentration, and size; this is not surprising, as these properties have been hypothesized to be important indications of the potential cytotoxicity of nanoparticles[98]. Very little work has been reported on the use of data mining and machine learning methods to predict the cytotoxicity of organic nanoparticles. One potential reason for this is the lack of databases or publications analyzing the cytotoxicity caused by a variety of organic nanoparticles. Another reason is the variability of biological models in different laboratories. Factors such as potential aggregation of nanoparticles, variations in the media used, cell origin and passage, among others, further contribute to variability in the

data obtained.

Only one of the articles reviewed[50] examined *in vivo* applications of data mining in nanomedicine. There is a clear lack of use of data mining and machine learning applications toward *in vivo* data regarding nanoparticles. Again this could probably be attributed to the lack of easily accessible data regarding *in vivo* applications of nanoparticles and the higher degree of variability of *in vivo* results.

Another commonality observed among many of the research articles presented in this review is the limited sample size and high dimensionality of the dataset used for analysis. Several consequences can arise due to lack of data, including overfitting; difficulty in demonstrating reliability, generalizability, and applicability of the predictive models to other nanoparticles; and class imbalance. Validation of a predictive model can be problematic when the sample size is limited and the variables representing those samples have high dimensionality[118]. The most common method for overcoming the issue of high dimensionality of a dataset is to utilize variable (feature) selection to reduce the number of variables analyzed in the predictive model[119]. Variable selection was commonly used in the research articles presented in this review, and as stated before, most of the researchers paired down their respective lists of variables to charge, concentration, and size-based properties of nanoparticles to create accurate predictions. Class imbalance is a challenging problem for the data mining community, it occurs when the samples representing one class are much lower than those representing other classes[120]. The simplest way to overcome this issue is to ensure that there is a balanced representation of the members of each class present in the dataset, but this is a significant challenge in nanoinformatics as the lack of large well-curated datasets seriously limits the amount, quality, and variety of data available. This is perhaps the most serious limitation observed in most of the papers discussed in this review. Since the field of nanoinformatics is relatively young, the data mining and machine learning results reported in the research

articles presented in this review are very preliminary and their generalizability is still an issue open for further investigation. As stated previously, it is our belief that NLP methods and the development of large curated databases with nanoparticle information will contribute to relieve the limitations commonly identified in most of the articles discussed here. Recently, review papers have focused on several challenges that face the development of nano-QSARs and other predictive models, including the lack of high-quality experimental data, lack of knowledge regarding interactions between nanoparticles like aggregation, high polydispersity in nanoparticles, etc.[121, 122]. These are definitely significant challenges that the field of nanoinformatics faces and should definitely be foci for future research.

The papers reviewed here clearly illustrate the power and accuracy data mining and machine learning methods bring toward creating predictions of functional and structural properties of nanoparticles. With the development of text mining, text extraction, and useful databases in the field of nanomedicine, the authors believe that the development of accurate nano-QSARs and other predictive models is quite possible with state of the art data mining and machine learning practices. Nonetheless, the great heterogeneity in nanoparticles will make these discoveries more challenging than for traditional small-molecule drug design.

Table 2.1: Summary of the systems studied, methods, and findings from papers using data mining and machine learning to predict cytotoxicity of nanoparticles. The evaluation results correspond to the most successful model reported in the publication.

First Author	Nanoparticle Type	Cell Type	Cytotoxicity Analysis	Predictive Method	Accuracy
Sayes [76]	Metal oxide nanoparticles (TiO_2)	Rat lung alveolar macrophages and immortalized rat L2 lung epithelial cells	LDH Release	LDA classification	$R^2 = 0.77$
Puzyn [60]	Metal oxide nanoparticles (ZnO, CuO, V_2O_3, Y_2O_3, Bi_2O_3, In_2O_3, Sb_2O_3, Al_2O_3, Fe_2O_3, SiO_2, ZrO_2, SnO_2, TiO_2, CoO, NiO, Cr_2O_3, and La_2O_3)	*E. coli*	EC_{50}	Multiple regression method	$R^2 = 0.85$
Liu [79]	Metal oxide nanoparticles (Al_2O_3, CeO_2, Co_3O_4, TiO_2, ZnO, CuO, SiO_2, Fe_3O_4, and WO_3)	BEAS-2B	Plasma membrane integrity	Logistic regression models	Accuracy = 100%
Horev-Azaria [80]	Metal oxide nanoparticle ($CoFe_2O_4$)	A549, NCI H441, HepG2, MDCK, Caco-2 TC7, TK6, and primary mouse dendritic-cells	Cell viability	J48	Accuracy = 92.5%
Winkler [67]	Metal oxide nanoparticle (CLIO)	Endothelial muscle cells, smooth muscle cells, hepatocytes, and monocytes	Smooth Muscle Apoptosis	Bayesian neural networks	$R^2 = 0.90$
Toropova [82]	Metal oxide nanoparticles	*E. coli*	pLC50	Monte Carlo Method	$R^2 = 0.9835$
Fourches [33]	Metal nanoparticles (CLIO, pseudo caged, monocrystalline iron oxide, CdSe core quantum dot, and iron-based)	Monocytes, hepatocytes, endothelial cells, and smooth muscle cells	Biological activity profiles	Support vector machine-based classification	Accuracy = 88%
Liu [50]	Metal nanoparticles, dendrimer, metal oxide, and polymeric materials	Embryonic zebrafish	24 hour post-fertilization mortality	IBK	Accuracy = 83.7%
Jones [74]	PAMAM dendrimers	Caco-2	Cell viability	J48	Accuracy = 83.5%

Table 2.2: Accuracies reported for models predicting different measurements of toxicity on zebra fish postfertilization embryos using different methods[50].

Predicted Attribute	Algorithm	Accuracy
120 hour postfertilization jaw malformation	IBK	0.667
120 hour postfertilization trunk malformation	IBK	0.657
24 hour postfertilization developmental progression	IBK	0.591
120 hour postfertilization pigmentation	IBK	0.565
120 hour postfertilization eye malformation	IBK	0.544
120 hour postfertilization snout malformation	IBK	0.486
120 hour postfertilization touch response	IBK	0.476
120 hour postfertilization caudal fin malformation	IBK	0.441
120 hour postfertilization yolk sac edema	Bagging	0.439
120 hour postfertilization pectoral fin malformation	IBK	0.387
120 hour postfertilization swim bladder	M5P	0.380
120 hour postfertilization circulation	IBK	0.368
120 hour postfertilization otic malformation	IBK	0.331
120 hour postfertilization brain malformation	IBK	0.297
120 hour postfertilization axis malformation	IBK	0.294
120 hour postfertilization somite malformation	Bagging	0.262
24 hour postfertilization notochord malformation	M5P	0.125
24 hour postfertilization spontaneous movement	KStar	-0.003

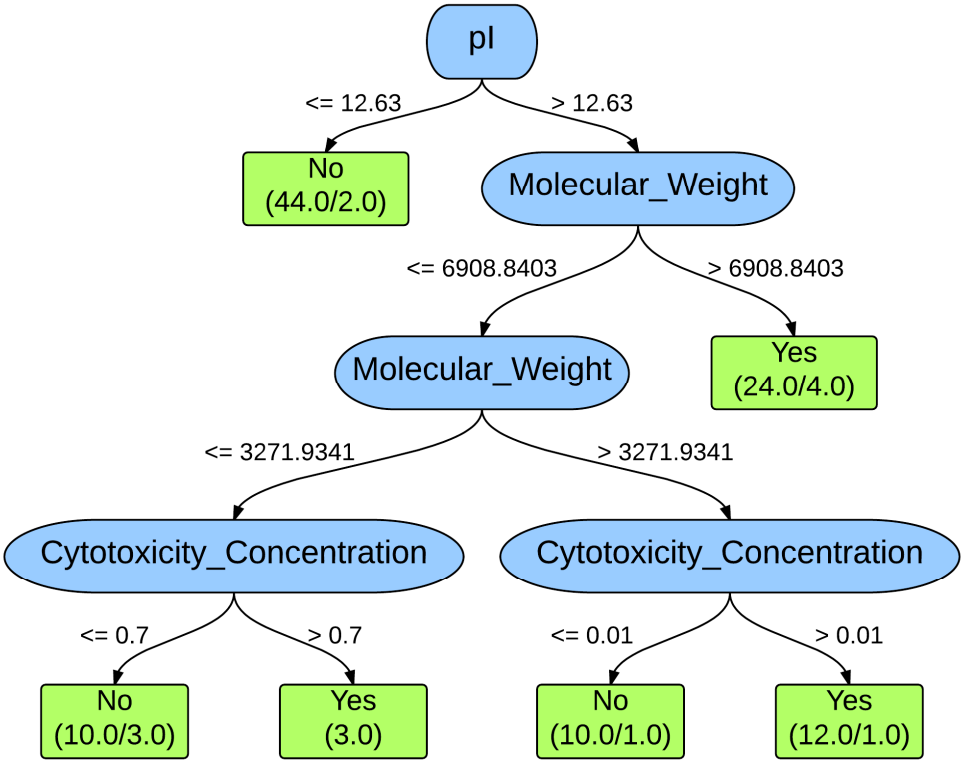

Figure 2.1: Decision tree for 10-fold cross-validation J48 classifier of the fourth analysis, including the molecular descriptors expert selected with the concentration information of dendrimers used in the experiments. Values present on the branches represent the rule or decision used for making the classification. The boxes at the bottom represent the classifications, with the number of PAMAM dendrimers classified as such on the left and the number of exceptions (misclassifications on the right). From Ref.[74] with permission.

Table 2.3: Summary of the data mining and machine learning methods used to predict molecular loading of nanoparticles. The evaluation results correspond to the most successful model reported in the publication.

Primary Author	**Nanoparticle Type**	**Loaded Molecule Type**	**Loading Type**	**Predictive Method**	**Evaluation Results**
Winkler [67]	Surface-modified gold nanoparticles	Protein	Nonspecific protein binding	Bayesian neural networks	$R^2 = 0.94$
Shalaby [99]	Di- and triblock copolymers of polyethylene glycol and polylactide	Noscapine	Entrapment efficiency	ANN	$R^2 = 0.96484$

Table 2.4: Summary of the data mining and machine learning methods used to predict molecular release from nanoparticles. The evaluation results correspond to the most successful model reported in the publication.

Primary Author	Nanoparticle Type	Released Molecule Type	Predictive Method	Evaluation Results
Husseini [101]	Polymeric (Pluronic P105) micelles	Doxorubicin	ANN	Maximum prediction errors = 0.001
Szlek [104]	Poly(lactic-co-glycolic acid) (PLGA) nanoparticles	Bovine serum albumin, human serum albumin, recombinant human erythropoietin, lysozyme, recombinant human epidermal growth factor, recombinant human growth hormone, beta-amyloid, recombinant human erythropoietin coupled with human serum albumin, hen ovalbumin, insulin, bovine insulin, L-asparaginase, chymotrypsin, and alpha-1 antitrypsin	ANN	RMSE = 15.4

Table 2.5: Summary of the data mining and machine learning methods used to predict nanoparticle size. The evaluation results correspond to the most successful model reported in the publication.

Primary Author	Nanoparticle Type	Predictive Method	Evaluation Results
Asadi [111]	Poly(lactide)-poly(ethylene glycol)-poly(lactide) (PLA-PEG-PLA) nanoparticles	ANN	$R^2 = 0.9434$
Shalaby [99]	Di- and triblock copolymers of poly(ethylene glycol) and poly(lactide)	ANN	$R^2 = 0.97833$

CHAPTER 3

AUTOMATIC EXTRACTION OF NANOPARTICLE PROPERTIES USING NATURAL LANGUAGE PROCESSING: NANOSIFTER AN APPLICATION TO ACQUIRE PAMAM DENDRIMER PROPERTIES[2]

Abstract

In this study, we demonstrate the use of natural language processing methods to extract from nanomedicine literature numeric values of biomedical property terms of poly(amido amine) dendrimers. We have developed a method for extracting these values for properties taken from the NanoParticle Ontology, using the General Architecture for Text Engineering and a Nearly-New Information Extraction System. We also created a method for associating the identified numeric values with their corresponding dendrimer properties, called NanoSifter.

We demonstrate that our system can correctly extract numeric values of dendrimer properties reported in the cancer treatment literature with high recall, precision, and f-measure. The micro-averaged recall was 0.99, precision was 0.84, and f-measure was 0.91. Similarly, the macro-averaged recall was 0.99, precision was 0.87, and f-measure was 0.92. To our knowledge, these are the first applications of text mining to extract and associate dendrimer property terms from their corresponding numeric values.

---

[2] *Reprinted from PLoS ONE, 9(1), Copyright 2014. David E. Jones [a], Sean Igo [a, b], John Hurdle [a], and Julio C. Facelli [a, b]. [a] Department of Biomedical Informatics and [b] Center for High Performance Computing, University of Utah, Salt Lake City, UT, US*

Introduction

Nanomedicine is the field of study that considers the application of nanoparticles and nanoscience techniques to health care and medical research[16]. A main focus of nanomedicine includes the use of nanoparticles as delivery vectors for pharmaceutics, diagnostic devices, and tissue replacement materials[2]. This field is relatively new, however it is producing large numbers of publications and substantial new data each year[1]. Data being published contain valuable information regarding how the structure of these nanoparticles relates to their biochemical and biophysical properties, which include but are not limited to their diameter, molecular weight, surface charge, zeta potential, bioavailability, cytotoxicity, etc.[5].

We have chosen dendrimers for our initial application of natural language processing (NLP) to nanomedicine because they are well-defined, highly branched polymeric nanoparticles that can easily be modified to differing specifications. There is also a substantial literature reporting their biological, chemical, and physical properties. Dendrimers are composed of a central core that is surrounded by concentric shells[44, 45]. The number of shells that extend out from the central core determines the particular generation of the dendrimer. Due to their structure, these molecules form very symmetric, three-dimensional particles that can be useful in the fields of pharmaceutics and medicine as delivery vectors[41]. The scaffold structure of dendrimers has been found to be a suitable carrier for a variety of drugs and siRNA, improving the solubility and bioavailability of poorly soluble agents. Currently there are several classes of dendrimers in use or under consideration for biomedical applications. This study focused on poly(amido amine) (PAMAM) dendrimers that show promise in cancer treatment.

Databases and repositories containing information relevant to biomedical nanoparticles, especially their biochemical and biophysical properties, are critical for both primary research as well as secondary uses such as data mining and predictive modeling.

The American National Standards Institute's Nanotechnology Standards Panel (ANSI-NSP) has created a Nanotechnology Standards database which is free for individuals and groups seeking information about standards and other relevant documents related to nanomaterials and nanotechnology-related products and processes[123]. The database does not directly host standards and other similar documents, but it provides a place for standards-developing organizations to add their relevant documents. This may someday be an important resource for the future development of standardized terminology in the field of nanotechnology and nanomedicine, but it does not contain an extensive collection of values of biological properties of medical nanomaterials.

nanoHUB.org is the premier site for computational nanotechnology research, education, and collaboration[124]. This resource provides an environment for collaboration and aggregation of tools used in simulating nanoscale phenomena. But with this resource the researchers must provide their own nanomaterial-specific data to utilize the host of simulation tools provided. To our knowledge, there is no authoritative, up-to-date database where researchers consistently contribute results from new publications on biomedical nanoparticles and their properties. Some attempts have been reported in the literature, like caNanoLab, a database created by the National Cancer Institute for sharing nanoparticle information[125]. However, caNanoLab contains a limited number of nanoparticles, and for those it often has incomplete information regarding their biological, chemical, and physical properties. Also, there are only limited capabilities to query this system. No data model exists to support comparing the properties of a molecule to its biochemical and biophysical activity. These properties are necessary to advance research on nanoparticles, but the only way to retrieve this information currently is by manual extraction from the primary literature.

Though manual extraction is a very time consuming and resource intensive process, little research has been done to apply computational methods to obtain

nanoparticle property data from the vast biomedical literature on nanoparticles. Information extraction (IE) efforts are widely acknowledged to be important in harnessing the rapid advance of biomedical knowledge, particularly in areas where important factual information is published in a diverse literature[27]. In particular, NLP is a family of methods based on syntactic/semantic analysis that can extract information automatically from the literature[28].

NLP has been used effectively in other biomedical domains. For instance, Chaussabel utilized NLP algorithms to extract data from the literature on cell line profiling. He observed that this approach could be applied beyond genomic data analysis[29]. Garten *et al.* successfully applied NLP methods to the pharmacogenomics literature to create structured databases built on data from unstructured text[30]. Hunter *et al.* created a system called OpenDMAP that extracts protein transport, interaction, and gene expression assertions[27]. In the field of nanoinformatics there has been an attempt at harnessing the utility of NLP in the nanomedicine literature by Garcia-Remesal and colleagues. They developed a method utilizing named entity recognition to identify four different categories of information: nanoparticle names, routes of exposure, toxic effects, and particle targets [126]. The method that this group developed was moderately successful, but it was designed as a proof-of-concept with limited quantitative detail. Our goal is to gather detailed quantitative data associated with dendrimer properties.

In this study, we evaluate the use of NLP methods to extract numeric values for the properties of biomedical dendrimers reported in the cancer treatment literature. We use open source tools for extracting particle property values, using the NanoParticle Ontology (NPO) [5] as a starting point. In particular, the tools we use are a processing pipeline called the General Architecture for Text Engineering (GATE) and its IE module ANNIE (a Nearly-New Information Extraction System) [127]. In a real-world sentence, a nanoparticle property term can appear arbitrarily far from its associated value, so we also created a

method of associating the two. We demonstrate that our system can correctly extract dendrimer property terms and their corresponding numeric values as evaluated by the typical NLP metrics of recall, precision, and f-measure score.

Materials and Methods

Literature Corpus

We collected from PubMedCentral relevant articles on dendrimer nanoparticles as reported in the cancer treatment literature. Articles were retrieved in pdf format. The search criteria used was "PAMAM dendrimers AND cancer treatment." This search yielded 420 journal articles on March 4, 2013. Articles were excluded from this study if they did not contain explicit numeric values of biological, chemical, and/or physical properties of dendrimers. From this pool of 420 articles, we randomly selected 200 journal articles. A subset of 100 articles was used as the training set for our system. The other subset of 100 articles was used for the creation of the test set for our system. Citations for both the training and test set of documents can be found in the supplementary information. For similar applications in related fields, the selection of a test set of approximately 100 documents is a common target that represents a compromise of quality and cost of the manual review. For instance Zaremba *et al.* used a test set of 138 abstracts to analyze enteropathogenic bacteria, such as *Escherichia coli* and *Salmonella*, literature [128].

NLP Method Development

The NLP system reported here uses a two-step process to extract the desired property terms and numeric values. The first step involves the actual identification and annotation of the numeric values and dendrimer property terms. This corpus annotation pipeline was built using the Java Annotations Patterns Engine (JAPE) and integrating components from ANNIE within GATE. In order to search for the numeric values, we had to develop a regular expression model (available in the supplemental materials). The

specific dendrimer property terms were selected from the NPO and represent the properties of nanoparticles. The dendrimer property terms were selected from the NPO with the ultimate goal of linking the NPO with our tool to provide metadata for the data extractions from the nanomedicine literature. The initial nanoparticle property terms list was confirmed to be relevant for the nanomedicine community by expert review by the members of Dr. Hamidreza Ghandehari's research lab (http://nanoinstitute.utah.edu/research/ustar-clusters/ghandehari-lab/ghandehari-PI.php) at the University of Utah. The list of terms considered here includes hydrodynamic diameter (NPO_1915), particle diameter (NPO_1539), molecular weight (NPO_1171), zeta potential (NPO_1302), cytotoxicity (NPO_1340), IC50 (NPO_1195), cell viability (NPO_1343), encapsulation efficiency (NPO_1336), loading efficiency (NPO_1334), and transfection efficiency (NPO_1335). The property terms, their corresponding NPO identification code, and their definitions can be found in Table 3.1. To search for these property terms, the system utilizes a simple keyword identification scheme.

The training set of documents was manually annotated for numeric values and dendrimer property terms using GATE. Following the annotation, the numeric values associated with each property term were extracted manually and organized in a tabular format for ease of use and comparison. Once the pipeline was able to successfully annotate the numeric values and the dendrimer property terms, we developed an algorithm that would associate numeric values and dendrimer property terms that occurred within the same sentence using proximity metrics. We selected a proximity distance metric of 200 characters because our preliminary experiments have shown that the sensitivity and specificity of the system was best for this distance in the training set. For instance we observed that if we increased it, the number of false positives increased without any improvement in the observed recall of the system. Finally, we optimized performance iteratively before moving on to the test set of documents.

Reference Standard Creation

Two domain experts were selected from the nanotechnology program at the University of Utah. Before allowing them to review the test subset of 100 articles, they independently reviewed, annotated, and extracted information from the training set of articles using GATE. The annotations consisted of numeric values and dendrimer property terms selected from the NPO. Their annotations were compared and Cohen's kappa was calculated. Cohen's kappa is a statistical measure of inter-rater reliability and for this study we required it to be ≥ 80%, which has been categorized as excellent by Fleiss at a value of 75% or higher [129].

Upon achieving an inter-rater reliability of 80%, the annotators independently reviewed, annotated, and extracted information from the test set of articles. Again, the numeric values and dendrimer property terms were taken from the NPO and were annotated using GATE. Following the annotation, the numeric values associated with each property term were extracted and organized in a tabular format.

NLP System Performance

The subset of 100 test articles was processed by our new NLP system. The output from the system was organized in a tabular format for ease of use and comparison.

Data Analysis

Our NLP and manual results were compared on a by-nanoparticle property term basis. The extracted numeric values associated to the dendrimer property terms were evaluated and determined to be true positive, false positive, or false negative. First, we calculated the recall, precision, and f-measure of each nanoparticle property term. We then calculated the micro-averaged and macro-averaged recall, precision, and f-measure. When using micro-averaged measurements, each "source" (e.g. document) is given the same weight and calculations are made on a pooled contingency table [130]. Macro-averaged

measurements are calculated by giving the same weight to each concept category or class (e.g., dendrimer property term) [130].

The recall, precision, and f-measure were calculated using the following equations:

$$Recall = TP / (TP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$F\text{-}measure = ((1 + \beta^2) * Precision * Recall) / ((\beta^2 * Precision) + Recall) \quad (3)$$

In these equations TP is true positive, FP is false positive, FN is false negative, and β is the weighting applied to the relationship between precision and recall. For our purposes we decided to weight the precision and recall evenly, so β=1.

Results

Table 3.2 summarizes the results of the evaluation of the NLP system that we created. The results of the system are compared against the manually annotated reference standard. The table shows the recall, precision, and f-measure for each of the nanoparticle property term and numeric value relationships. Table 3.3 displays both the micro-averaged and macro-averaged recall, precision, and f-measure values.

As can be seen in Table 3.2, our NLP system yields recall values ranging from 0.95 to 1, precision values range from 0.59 to 1. The f-measure values range from 0.73 to 1. The micro-averaged values for recall was 0.99, precision was 0.84, and f-measure was 0.91. Similarly, the macro-averaged values for recall was 0.99, precision was 0.87, and f-measure was 0.92.

Discussion

The tables show an important difference between recall and precision. In this task, high recall is preferred to high precision because we do not want our system to miss instances of property terms and their associated numeric values. The number of articles returned for any given search (e.g., our "PAMAM dendrimers AND cancer treatment" search) is too large for routine manual search, but reviewing NanoSifter *results* is quite tractable. The results can be manually reviewed post-processing without much additional effort. From the results, it can be seen that "encapsulation efficiency" and "loading efficiency" were the best property terms extracted with recall, precision, and f-measure values of 1. These scores are likely due to the low prevalence of these properties appearing in our literature corpus. "Transfection efficiency" was the property term that was the least well extracted from nanomedicine literature. It had a recall value of 0.95, a precision value of 0.59, and an f-measure value of 0.73.

These results indicate that the Nanosifter NLP system can, generally, extract numeric values associated with particle property terms from dendrimers reported in the cancer treatment literature with high recall, precision, and f-measure scores. To the authors' knowledge, these results are the first application of text mining to extract numeric values associated to dendrimer property terms from nanomedicine literature. With regards to our application, the high recall values are more important than the moderate precision values. This is because the lack of precision is manageable and can be quickly corrected by manual post processing of the annotated text.

As can be seen from the results, there was a fair amount of fluctuation in the values for precision for each property term. There were a few property terms that yielded precisions of 1 including "hydrodynamic diameter," "zeta potential," "encapsulation efficiency," and "loading efficiency." This can be accounted for by the limited number of instances that these terms appeared in the literature. Of all of the property terms used in

this study, these were the least common. The next tier of precision values of interest are those that were greater than 0.80, these include “particle diameter,” “molecular weight,” “cytotoxicity,” and “IC50.” These property terms yielded quite reasonable precision values, as we expected based upon their occurrences in the literature and the specificity of the syntax used when describing these property terms and their numeric values.

The lowest precision values could be seen for “cell viability” (0.72) and “transfection efficiency” (0.59). One reason for these lower precision values is that the numeric units for these properties are percentages. There was a significant number of false positives in the literature corpus because the number of occurrences of percentages for other, non-particle items within the 200-character proximity metric was large. With specific regard to “transfection efficiency,” precision values for this term were the lowest because the terminology used to refer to this property is not standardized. There are many different ways in which the literature refers to this property, making it difficult not to overfit a method of retrieving the numeric values of this property.

Limitations

NanoSifter uses a method that appears to be generally reliable and accurate. However, there are imperfections that were observed while processing and analyzing the data from this study. First, the data extracted by our method is not always directly associated with a dendrimer nanoparticle. For instance, many times the system correctly finds, annotates, and extracts a “molecular weight measurement”, but this measurement may be associated with a subunit utilized in the synthesis of a PAMAM dendrimer or another material used in one of the articles. A method to address this limitation could include post-analysis manual review of the system’s performance. Another limitation of our system is that the NanoSifter algorithm can only pair a nanoparticle property term with a single numeric value annotation before and after itself. This causes a problem when

a sentence is more complex and contains a property term, random text, numeric value, random text, or another numeric value. In NLP, this is a problem called co-reference resolution, and it could be addressed with a more sophisticated language model than the one used in this study.

Another limitation is that our system would only retrieve the first numeric value expressed following the property term. This situation accounts for some of the false negatives ("particle diameter," "cytotoxicity," "IC50," and "transfection efficiency") found in our analysis. This could also be addressed by using a more sophisticated language model than the one used in this study. Finally, the other false negatives, "molecular weight" and "zeta potential," account for another limitation of our system. Since we were processing pdf documents in this study, occasionally there would be an instance where a property term exceeded a single line of text, so a dash would be inserted in the word and it would continue on the next line. The method used in developing this system did not account for this artifact, so the NanoSifter NLP system would not annotate this property term and no association would be made to the corresponding numeric value. A method for addressing this would be to use XML documents instead of pdfs in future analyses. These limitations are not novel to our approach, as they are common throughout the field of NLP. Nonetheless they are counterbalanced by the ability to extract information from journal articles at a much lower cost than manual review.

Future Work

Since this is early work in an important but neglected area of nanoinformatics there are many directions this research could be taken. The first priority will be to make corrections to our system to try to improve our recall, precision, and f-measure values. Another priority will be to attempt to use this system to annotate and extract information from another subclass of nanoparticles. This will help to validate the ability of this system

to generalize across the field of nanoparticles. One of the most important next steps would be to expand the property terms and numeric values that the system targets. Some specific properties that we are considering include “exposure times” and “cell types” interacting with the nanoparticles. This would allow for greater databases to be created regarding PAMAM dendrimers and nanoparticles in general. Another goal would be to more seamlessly integrate the NPO into our system so that the annotations and extractions contain descriptive metadata. Finally, it is important that we attempt to implement some sort of negation analysis tool into our system. This would specifically help in the instances where an article states that the dendrimer nanoparticles were not toxic at a certain concentration.

Conclusion

In this paper, we have presented a nanoinformatics method based on NLP approaches for automatically extracting numeric values associated with dendrimer property terms from the nanomedicine literature. The results from our analysis demonstrate that the NanoSifter NLP system can be used to reliably and accurately extract information from dendrimers developed for cancer treatment literature and shows promise for the future of text mining in the field of nanoinformatics. This initial research in the field of applying NLP to nanomedicine literature could assist in significant advances for the nanomedicine community. This work could lead to the creation of databases containing valuable information regarding nanoparticles at a much lower cost than using manual review. The readily available data on nanomedical relevant particles could be further analyzed for many secondary uses of the data. In particular, the acquired data could be used for data mining to find correlations between properties, create predictive models like quantitative structure activity relationships, and eventually reach the point where potential candidate molecules can be created *in silico* and modeled to theoretically

predict their biochemical activity before synthesis. This would reduce the search space for novel, effective nanoparticles for use in medicine and pharmaceutics.

Table 3.1: Listing of the NPO Property Terms

PROPERTY TERM	NPO CODE	DEFINITION
Hydrodynamic Diameter	NPO_1915	The hydrodynamic size which is the diameter of a particle or molecule (approximated as a sphere) in an aqueous solution.
Particle Diameter	NPO_1539	Diameter which inheres in a particle.
Molecular Weight	NPO_1171	The sum of the relative atomic masses of the constituent atoms of a molecule.
Zeta Potential	NPO_1302	The potential difference between the bulk dispersion medium (liquid) and the stationary layer of liquid near the surface of the dispersed particulate.
Cytotoxicity	NPO_1340	Toxicity that impairs or damages cells, and it is a desired property of the dispersed particulate.
IC50	NPO_1195	A measure of toxicity which is the concentration of a drug or inhibitor that is required to inhibit a biological process or a participant's activity in that process by half.
Cell Viability	NPO_1343	Viability of a cell to proliferate, grow, divide, or repair damaged cell components.
Encapsulation Efficiency	NPO_1336	The efficiency of inhering in a nanomaterial or supramolecular structure by virtue of its capacity to encapsulate an amount of molecular entity, isotope or nanomaterial.
Loading Efficiency	NPO_1334	A quality inhering in a material entity by virtue of it having the capacity to carry an amount of another material entity.
Transfection Efficiency	NPO_1335	The efficiency inhering in a bearer's ability to facilitate transfection.

Table 3.2: Results from the Evaluation of the Nanosifter NLP System

Nanoparticle Property Term	**TP**	**FP**	**FN**	**Recall**	**Precision**	**F-measure**	**Occurrences by Article**
Hydrodynamic Diameter	8	0	0	1	1	1	6
Particle Diameter	211	39	1	0.995283	0.844	0.91341991	56
Molecular Weight	143	23	2	0.986207	0.86145	0.91961415	25
Zeta Potential	41	0	1	0.97619	1	0.98795181	16
Cytotoxicity	124	18	1	0.992	0.87324	0.92883895	29
IC50	47	8	1	0.979167	0.85455	0.91262136	15
Cell Viability	78	31	0	1	0.7156	0.8342246	25
Encapsulation Efficiency	1	0	0	1	1	1	1
Loading Efficiency	5	0	0	1	1	1	1
Transfection Efficiency	19	13	1	0.95	0.59375	0.73076923	9

Table 3.3: Micro-averaged and Macro-averaged Recall, Precision, and F-measure

Type of Average	**Recall**	**Precision**	**F-measure**
Micro	0.989766	0.83684	0.90689886
Macro	0.987885	0.87426	0.922744

CHAPTER 4

PREDICTING CYTOTOXICITY OF PAMAM DENDRIMERS USING MOLECULAR DESCRIPTORS[3]

Abstract

The use of data mining techniques in the field of nanomedicine has been very limited. In this paper we demonstrate that data mining techniques can be used for the development of predictive models of cytotoxicity of poly(amido amine) (PAMAM) dendrimers using their chemical and structural properties. We present predictive models developed using 103 PAMAM dendrimer cytotoxicity values that were extracted from 12 cancer nanomedicine journal articles. The results indicate that data mining and machine learning can be effectively used to predict cytotoxicity of PAMAM dendrimers on Caco-2 cells.

Introduction

In silico approaches, such as data mining and machine learning, have been very successful in medicinal chemistry and are commonly used to guide the design of small pharmaceutical compounds[20]. In contrast, even while nanomedicine is a rapidly growing field[5], there have been only a few attempts to use data mining techniques in this field. For instance, Liu *et al.* analyzed a number of attributes of a variety of nanoparticles in

---

[3] *Reprinted from Beilstein Journal of Nanotechnology, 6, pp 1886-1896, Copyright 2015. David E. Jones [a], Hamidreza Ghandehari [b, c], and Julio C. Facelli [a, b]. [a] Department of Biomedical Informatics, [b] Departments of Bioengineering and Pharmaceutics and Pharmaceutical Chemistry, and [c] Utah Center for Nanomedicine, Nano Institute of Utah, University of Utah, Salt Lake City, UT 84112, US*

order to predict the 24 hours postfertilization mortality in zebrafish[50]. Horev-Azaria and colleagues used predictive modeling to explore the effect of cobalt-ferrite nanoparticles on the viability of seven different cell lines[80]. Sayes and Ivanov used machine learning to predict the induced cellular membrane damage of immortalized human lung epithelial cells caused by metal oxide nanomaterials[76]. As discussed in a previous paper[131], there are very limited databases of properties of nanomedical relevant compounds. We speculate that this has seriously limited the use of data mining techniques in the field of nanomedicine, but in the above referenced publication we demonstrated that natural language processing (NLP) techniques can be used effectively to automatically extract nanoparticle property information from the original literature. Here we argued that this development opens the possibility to explore the use of data mining and chemometric techniques to guide the design of new, more effective, treatments using nanoparticles. In this paper we apply the methods of data mining and machine learning to predict the cytotoxicity of poly(amido amine) (PAMAM) dendrimers.

Cytotoxicity was selected because it is of key concern for the nanoscience and nanomedicine community[7, 8], considering that high cytotoxicity is a definitive cause for eliminating a material for potential human applications. Reliable prediction of cytotoxicity using *in silico* approaches possesses the potential for high payoff in nanomaterial development, allowing the concentration of scarce development resources to be directed towards the synthesis and testing of promising materials with expected low levels of toxicity. Cytotoxicity can be determined by a gamete of *in vitro* toxicity assays focusing on a number of cellular parameters including cell viability, oxidative stress, genotoxicity, and inflammatory response[75]. In this paper, we focus on the cell viability to characterize cytotoxicity[18].

PAMAM dendrimers are good candidates for a data mining methodological study because they are well documented and have the potential to be highly useful as delivery

vectors[41]. These nanoparticles are composed of a central core that is surrounded by concentric shells, thus resulting in their well-defined, highly branched structure[42, 43]. The generation of the dendrimer is determined by the number of concentric shells that surround the core of the structure. These polymeric nanoparticles can easily be tailored for specific applications. Benefiting from their characteristic scaffold structures, they have been demonstrated to be suitable carriers for a number of diverse bioactive agents, improving the solubility and bioavailability of poorly soluble ones[44, 45]. These particular nanoparticles are also promising for use in the treatment of cancer, including oral formulations. In spite of all the desirable properties of dendrimers, there is a significant setback for their use in biomedicine due to their potential toxicological effects, which depend on the structure that is used. It has been shown that cationic PAMAM dendrimers can have surface charge-, generation-, and concentration-dependent toxicity[46-49].

The goal of this research is to demonstrate that data mining methods like the ones used here can provide a presynthesis step to identify nondesirable PAMAM dendrimers that have a substantial probability of high toxicity. It would thus be possible to eliminate them from the early stages of the synthetic development pipeline with reasonable confidence. This technique is not meant to replace cytotoxicity assays at the bench, but rather to augment these methods. This method will bolster existing cytotoxicity assays by providing the ability to determine relevant compounds with low cytotoxicity and to eliminate weak-candidate PAMAM dendrimers from synthesis and confirmatory testing. This work also illustrates a proof of concept that data mining and machine learning can be applied to PAMAM dendrimers to predict their biochemical properties. This result could potentially be expanded to other nanomaterials in the future.

Results and Discussion

Five different analyses were performed to classify a dendrimer as toxic or nontoxic using different combinations of molecular descriptors and experimental conditions. The first analysis utilized all the molecular descriptors available in MarvinSketch (see Experimental section and the Appendix). The second analysis involved an automatic feature selection method in which the molecular descriptors that were used had a nonzero rank according to the ChiSquaredAttributeEval method in Weka (see details in the Experimental section). The ChiSquaredAttributeEval method determines the rank of an attribute by calculating the chi-squared statistic with respect to the class[92]. The third analysis used only the molecular descriptors selected by expert advice (see details in the Experimental section): molecular weight, atom count, pI, and molecular polarizability. The fourth analysis included the same molecular descriptors used in the second analysis and the experimental concentration (i.e., the amount in mM of PAMAM dendrimer added to the human colon carcinoma Caco-2 cells culture during the cytotoxicity analysis). The final analysis independently assessed the performance of our best method by randomly splitting the dataset into a training set, including 83 of the values, and a test set, including 20 of the values in the dataset.

The results for the first, second, and third analyses performed to classify dendrimers as toxic/nontoxic are presented in Tables 4.1-4.3 and the Appendix. The tables list the average precision, recall, F-measure, and mean absolute error for the toxicity class prediction for all classifiers considered here. The tables also contain the accuracy value for the percentage of correctly classified instances. For all analyses all classifiers consistently have accuracies at or above 60.2%.

For the first analysis, Tables 4.1 and the Appendix, J48 and filtered classifier show the best results in the 10-fold cross-validation with an accuracy of 74.8%, while bagging, locally weighted learning (LWL), and naïve Bayes Tree (NBTree) performed the best with

an accuracy of 77.7% in the leave-one-out cross-validation (Appendix). The results from the automatic feature selection analysis, using the ChiSquaredAttributeEval and Ranker procedures as attribute evaluator and search method, respectively, are presented in Table 4.2 and the Appendix. These results do not differ drastically from those observed in the first analysis, indicating that the use of automatic feature selection does not improve the classification of toxicity in this study. Alternative automatic feature selection methods, using all the WEKA recommended pairings of attribute evaluator and search methods were also tested but did not show any significant improvement in classification prediction performance when using the J48 classifier. These results are presented in the Appendix. The classification using the features selected by expert advice, Table 4.3 and the Appendix, show that the LWL classifier performed the best with an accuracy of 77.7% in the 10-fold cross-validation. The leave-one-out cross-validation, in the Appendix, had three classifiers that perform with an accuracy of 78.6% (naïve Bayes, bagging, and classification via regression). There is an increase in accuracy across most of the classifiers between the 10-fold and leave-one-out cross-validations. This is an interesting finding because Kohavi noted that k-fold cross validations typically perform better than leave-one-out cross validations[132]. This might be an artifact of the dataset not being exactly 50-50 split between toxic and nontoxic samples, thus leading to skewness toward nontoxic predictions.

The decision tree used by the 10-fold and leave-one-out cross-validation J48 classifiers for the first, second, and third analyses is depicted in Figure 4.1. It can be observed in the decision trees that the isoelectric point, pI, is the property that is used to classify the dataset. This property represents the pH at which the net charge of an ionizable molecule is zero. The decision tree indicates that if the pI is greater than 12.63, then the dendrimers are toxic. There are 59 PAMAM dendrimers that are classified as toxic of which 21 are misclassified. If the pI is less than or equal to 12.63, then the dendrimers are classified as nontoxic. There are 44 PAMAM dendrimers classified as nontoxic of which 2

are misclassified.

These results indicate that data mining and machine learning can be implemented to predict cytotoxicity of PAMAM dendrimers on Caco-2 cells with reasonably high accuracy using only molecular descriptors. The misclassifications observed in Figure 4.1 are much more significant when examining the dendrimers classified as toxic because almost half of these dendrimers are nontoxic. This is a substantial quantity of potentially useful dendrimers being ruled out, indicating the necessity for further analysis to decrease the number of false positives.

Table 4.4 presents the results using the best performing classifiers from the previous section of the analysis using the expert-selected molecular descriptors with the addition of the concentration of dendrimers used in the experiments. No improvement in predictions was observed when using either the Filtered Classifier or LWL classifiers, but the J48 prediction accuracy of the classification improved to 83.5%. This substantial improvement of the accuracy of the J48 classifications, from 74% to 83.5 %, shows the importance of including the concentration information from the experimental design in addition to the computed molecular descriptors to properly classify compounds as toxic or nontoxic.

The J48 decision tree for the analysis discussed above is depicted in Figure 2.1. In this case, the pI, molecular weight, and cytotoxicity concentration are the discriminators in the classification. As can be seen, the feature representing the concentration of dendrimers used in the experiments is present in the decision tree for this analysis. The diagram of the decision trees generated from the J48 classifier illustrates important attributes used in accurately predicting toxicity in PAMAM dendrimers. The greatest prediction accuracies were achieved after supplementing the expert-selected features with a descriptor representing the experimental conditions by including the concentration under which the data of the cytotoxicity has been acquired. Figure 2.1 has the same

structure at the top level as Figure 4.1: when the pI is less than or equal to 12.63, 44 PAMAM dendrimers are classified as nontoxic with an exception of 2 that are misclassified. However, when the pI is greater than 12.63, it leads to other options in the classification of the remaining PAMAM dendrimers. The decision made at the next node regards the molecular weight of the PAMAM dendrimer being <= 6908.8 Da or > 6908.8 Da. If the molecular weight is > 6908.8 Da, 24 PAMAM dendrimers are classified as toxic with 4 that are misclassified. If the molecular weight is <= 6908.8 Da, it leads to another option regarding the molecular weight being <= 3271.9 Da or > 3271.9 Da. The final option can be made considering the concentration target for the desired application of the PAMAM dendrimer. It can clearly be observed in Figure 2.1 that the number of misclassifications (false positives) has been significantly reduced due to this further analysis (from 21 in Figure 4.1 to 5 in Figure 2.1). Due to the significant decrease in false positives, the accuracy of the J48 classifier improved. There was a slight increase in the number of false negatives due to this further analysis (from 2 in Figure 4.1 to 5 in Figure 2.1).

The classification scheme in Figure 2.1 identifies three clusters of viable PAMAM dendrimers that have tolerable levels of cytotoxicity: those with a pI less than or equal to 12.63; those with a pI greater than 12.63, but with molecular weights less than or equal to 3271.9 Da that could be used up to concentrations of less than or equal to 0.7 mM; and those with a pI greater than 12.63, with molecular weights between 6908.8 and 3271.9341 Da that can be used in formulations requiring concentrations less than or equal to 0.01 mM. When designing novel PAMAM dendrimers, these guidelines could be used for developing viable candidates exhibiting little to no cytotoxicity. This demonstrates the importance of combining experimental conditions with molecular descriptors to achieve the greatest prediction accuracy in the classifiers and to find compounds that may be viable under more restrictive conditions. Another important observation is that the

properties present in the decision tree diagrams represent the more general properties of charge, size, and concentration, which have been hypothesized to be primary causes of cytotoxicity in Caco-2 cells [98].

Tables 4.5 and 4.6 show the data from the external validation study that was performed to further validate the results presented above. For this study, the dataset was randomly split into a training set, consisting of 83 cytotoxicity values, and a test set, consisting of 20 cytotoxicity values from the original dataset. Table 4.5 presents the results from the analysis of this test set using all of the molecular descriptors. For all but one of the classifiers, the predicted accuracy was 65.0%, which is slightly lower than the values obtained for the cross validation analysis, but LWL performed very well with an accuracy of 95.0%. This is an interesting finding considering the highest performance of this classifier in the first four analyses was 77.7%. Table 4.6 shows the data from the analysis of the test set using only the expert-selected features as well as the cytotoxicity concentration data. Again, LWL performed with an accuracy of 95.0%, so no improvement was seen in the classification ability of this algorithm between all molecular descriptors and the expert-feature-selected molecular descriptors with cytotoxicity concentration data. There are two algorithms that exhibited a large improvement between Tables 4.5 and 4.6, naïve Bayes and J48. Both of these algorithms improved from a prediction accuracy of 65.0% to 90.0%, which is substantially higher than the values obtained in the cross validation studies.

These results indicate that data mining and machine learning can be implemented to predict cytotoxicity of PAMAM dendrimers on Caco-2 cells with accuracy. According to Figure 2.1, the results also indicate that properties regarding charge, size, and the desired concentration of the PAMAM dendrimers in the formulation are the important properties in the prediction of cytotoxicity on Caco-2 cells. We believe that the methods used in this work can be expanded to analyze and predict many other biochemically relevant

properties of not only unmodified PAMAM dendrimers but also surface-modified PAMAM dendrimers. This method will bolster existing cytotoxicity assays by providing the ability to determine relevant compounds with low cytotoxicity for synthesis and confirmatory testing, thereby reducing the search space necessary for developing biomedically relevant PAMAM dendrimers. This work demonstrates a proof of concept that data mining and machine learning can be applied to PAMAM dendrimers to predict the biochemical property of cytotoxicity, but also indicates that further studies including much larger data sets are necessary to develop reliable and robust classification methods that can be applied to a broader set of compounds, cell cultures and experimental designs.

Conclusions

In this study, classification methods for predicting the Boolean classification of cytotoxicity in Caco-2 cells treated with PAMAM dendrimers were introduced. The results indicate that data mining and machine learning can be used to predict cytotoxicity of PAMAM dendrimers on Caco-2 cells with good accuracy. In the classification method explored here, it was observed that the properties regarding charge, size, and concentration of the PAMAM dendrimers are the most important properties in the prediction of cytotoxicity and cell viability of Caco-2 cells treated with PAMAM dendrimers. To the authors' knowledge, these results are the first application of data mining and machine learning to predict cytotoxicity of PAMAM dendrimers on Caco-2 cells using a classification method.

Experimental

The overall work flow of the analysis reported in this paper is presented in Figure 4.2. The details of the different processes are given in the following subsections.

Nanoparticle Selection

The PAMAM dendrimers selected for our study included generations 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, and 4.5 compounds that have been used for transepithelial transport. The full generation PAMAM dendrimers (generations 0, 1, 2, 3, and 4) are amine- or hydroxyl-terminated dendrimers. The half generation PAMAM dendrimers (generations 1.5, 2.5, 3.5, and 4.5) are carboxyl-terminated dendrimers. For more general property information on the full and half generation PAMAM dendrimers see the Appendix, which includes the property information for the PAMAM dendrimers analyzed in this study. The toxicity studies used here correspond to assays of these compounds on the human colon carcinoma Caco-2 cell line. The publications containing property data for the nanoparticles selected for this study were gathered from nanomedicine articles available in Scopus and PubMedCentral using the search terms "PAMAM dendrimers AND cytotoxicity AND Caco-2 cells." For the PAMAM dendrimer cytotoxicity values to be considered relevant for extraction, both cell viability and treatment concentration information had to be available in the publication. From this literature corpus, 103 PAMAM dendrimer cytotoxicity values were extracted to be included in this study[133-144]. NanoSifter[131], followed by manual revision, was used to extract the cell viability and cytotoxicity treatment concentration information from the journal articles in the corpus described above.

Chemical Structure Rendering and Molecular Descriptor Calculation

The PAMAM dendrimers' structures were manually constructed using MarvinSketch by ChemAxon[97, 145]. There were a total of 10 PAMAM dendrimer structures created for this study. They included generations 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, and 4.5 PAMAM dendrimers. These models include both amine-terminated (full generations) and carboxyl-terminated (half generations) structures, as well as one hydroxyl-terminated structure (full generation but hydroxyl-terminated). The molecular descriptors for each

molecule were calculated using plugins built into MarvinSketch[97]. The list of 51 molecular descriptors calculated for each molecule is given along with their corresponding definitions in the Appendix. Among these molecular descriptors, there are 42 structural properties (2 mass-related, 6 atom-count-related, 7 bond-count-related, 4 ring-size-related, 13 ring-count-related, and 10 other structural properties) and 9 chemical properties (5 charge-related and 4 hydrogen-bonding-related properties).

Data Preparation and Pre-Processing

The data, consisting of the molecular descriptors calculated for all of the molecules considered here and the corresponding cell viability and cytotoxicity data, was uploaded into WEKA [92] to perform the machine learning and data mining analysis using classification methods to discern between toxic and nontoxic compounds. In order to assign a categorical value to each dendrimer cytotoxicity data point, the threshold was established at the cell viability value of 90% (i.e., compounds were considered nontoxic at a certain concentration of PAMAM dendrimer nanoparticles if 90% of the Caco-2 cell population survived after the intervention). Because there is a statistical variation in cell viability studies, often non-toxic materials can have a few percentages above or below 100% cell viability. Hence, the threshold of 90% was set arbitrarily to take into account the usual variability in this type of study.

Prediction of Toxicity Using Classification Methods

Five different analyses were performed to classify a dendrimer as toxic or nontoxic using different combinations of molecular descriptors and experimental conditions. The first analysis utilized all the molecular descriptors. The second analysis involved an automatic feature selection using the ChiSquaredAttributeEval and Ranker method built into Weka, where only molecular descriptors with a nonzero rank were included in this analysis. The molecular descriptors with a nonzero rank were H-bond acceptor sites,

isoelectric point (pI), logP, Harary index, refractivity, bond count, molecular polarizability, rotatable bond count, atom count, logD, aliphatic bond count, chain bond count, chain atom count, aliphatic atom count, exact mass, molecular weight, Wiener index, Randic index, Szeged index, Wiener polarity, Platt index, H-bond donor count, hyper Wiener index, H-bond donor sites, and H-bond acceptor count. The third analysis used only molecular descriptors selected by expert advice: molecular weight, atom count, pI, and molecular polarizability. In this paper we refer to selected by expert advice as the properties that an experienced researcher in nanocarriers, Dr. Ghandehari, expected to be relevant to predict toxicity based on his own knowledge derived from work in his lab and literature precedents. The fourth analysis included the same molecular descriptors as the ones used in the second analysis and the experimental concentration, i.e., the amount in mM of PAMAM dendrimer added to the Caco-2 cells during cytotoxicity analysis. The fifth analysis was an external validation study, in which we randomly selected 20 cytotoxicity values from the original dataset of 103 to create a test set. The remaining 83 cytotoxicity values were used as the training set.

In this work we used the following classifiers: naïve Bayes, sequential minimal optimization (SMO), J48, bagging, classification via regression, filtered classifier, LWL, decision table, decision table/naïve Bayes (DTNB), NBTree, and random forest. We wanted to explore many modeling methods to provide a wide landscape of available techniques. Since the computational cost is low, there is no strong argument to limit this exploration. Naïve Bayes is a Bayesian classifier which uses posterior probability to predict the value of the target attribute[89]. That is, by using a given input attribute, the classifier attempts to find the target attribute value that maximizes the conditional probability of the target attribute. SMO is a support vector machine classifier that globally replaces all values and transforms nominal attributes into binary ones [90]. By default it normalizes all attributes. J48 is a decision tree classifier, which is based on the C4.5 algorithm[81]. This

method starts with large sets of cases which belong to known classes, then cases are analyzed for patterns that allow for reliable discrimination of classes. The patterns are represented as models, either in the form of decision trees or sets of if-then rules which can be used to classify new cases. Bagging is a hybrid classification method that creates classes and reduces variance by bagging classifiers[84]. Classification via regression performs its classification by binarizing each class and building one regression model for each class[91]. Filtered classifier is an arbitrary classifier that runs on data passed through an arbitrary filter[92]. LWL uses an instance-based algorithm to assign instance weights[93]. Decision table is a simple decision table majority classifier[94]. DTNB is a decision table/naïve Bayes hybrid classifier. During the search, the algorithm determines the need to divide the attributes into two disjoint subsets: one for the decision table, the other for naïve Bayes[95]. NBTree is a decision tree/naïve Bayes hybrid classifier that builds a decision tree with Naïve Bayes classifiers at the leaves[96]. All the calculations were performed using WEKA[92].

Two different cross-validation[146] schemes were performed for each classifier. The first one was a 10-fold cross-validation, in which the dataset was divided into 10 parts or folds[92]. During each classification run, nine of the folds were used as a training set and one was used as a test set and the results averaged over the ten runs. The second cross-validation scheme used here was the leave-one-out cross-validation[92]. As this cross-validation method states, one sample is left out as the test set, and the rest of the dataset is the training set. This method runs this through as many iterations as there are samples in the dataset.

The predictions determined by WEKA were evaluated and determined to be true positive, false positive, or false negative by manual inspection. The precision, recall, and F-measure were calculated using the following equations:

$$Precision = TP / (TP + FP) \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

$$F\text{-}measure = ((1 + \beta^2) * Precision * Recall) / ((\beta^2 * Precision) + Recall) \quad (3)$$

$$Mean\ Absolute\ Error = (\Sigma\ |f_i - y_i|)/n \quad (4)$$

In these equations, TP is true positive, FP is false positive, FN is false negative, and β is the weighting applied to the relationship between precision and recall. The precision and recall were weighted evenly, so β=1 [6]. The precision, recall, and F-measure of each classifier were calculated for each classification (toxic/nontoxic). Each measure for each classification (toxic/nontoxic) was then averaged. The average value for the precision, recall, and F-measure were recorded. For mean absolute error, f_i is the prediction, y_i is the true value, and n is the number of calculated absolute errors.

Table 4.1: Results from the 10-fold cross-validation listed by classifier of the first analysis including all molecular descriptors. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.654	0.660	0.655	0.3370	66.0%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.746	0.738	0.740	0.3211	73.8%
Classification via Regression	0.734	0.738	0.730	0.2978	73.8%
Filtered Classifier	0.789	0.748	0.750	0.3077	74.8%
LWL	0.775	0.738	0.741	0.2966	73.8%
Decision Table	0.678	0.660	0.664	0.3878	66.0%
DTNB	0.691	0.670	0.674	0.3490	67.0%
NBTree	0.696	0.670	0.674	0.3511	67.0%
Random Forest	0.736	0.718	0.722	0.3077	71.8%

Table 4.2: Results from the 10-fold cross-validation listed by classifier of the second analysis including the automatically feature selected molecular descriptors. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.654	0.660	0.655	0.3370	66.0%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.746	0.738	0.740	0.3211	73.8%
Classification via Regression	0.734	0.738	0.730	0.2978	73.8%
Filtered Classifier	0.789	0.748	0.750	0.3077	74.8%
LWL	0.775	0.738	0.741	0.2966	73.8%
Decision Table	0.678	0.660	0.664	0.3878	66.0%
DTNB	0.691	0.670	0.674	0.3490	67.0%
NBTree	0.696	0.670	0.674	0.3572	67.0%
Random Forest	0.736	0.718	0.722	0.2988	71.8%

Table 4.3: Results from the 10-fold cross-validation listed by classifier for the third analysis including the molecular descriptors selected by experts. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.762	0.748	0.750	0.2822	74.8%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.789	0.748	0.750	0.3077	74.8%
Bagging	0.731	0.718	0.721	0.3217	71.8%
Classification via Regression	0.762	0.748	0.750	0.3230	74.8%
Filtered Classifier	0.804	0.757	0.760	0.3061	75.7%
LWL	0.834	0.777	0.778	0.3008	77.7%
Decision Table	0.658	0.650	0.653	0.3980	65.0%
DTNB	0.658	0.650	0.653	0.3969	65.0%
NBTree	0.722	0.689	0.693	0.3454	68.9%
Random Forest	0.758	0.748	0.750	0.2973	74.8%

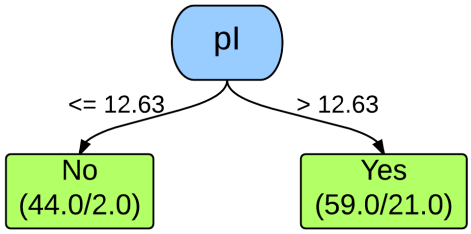

Figure 4.1: Decision tree for both 10-fold and leave-one-out cross-validation J48 classifier of the first, second, and third analyses. Values present on the branches represent the rule or decision used for making the classification. The boxes at the bottom represent the classifications with the number of PAMAM dendrimers classified as such on the left and the number of exceptions (misclassifications on the right).

Table 4.4: Results from the 10-fold cross-validation listed by classifier of the fourth analysis including the expert selected molecular descriptors with cytotoxicity concentration. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.755	0.738	0.741	0.2984	73.8%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.838	0.835	0.836	0.2203	83.5%
Bagging	0.836	0.835	0.835	0.2618	83.5%
Classification via Regression	0.742	0.738	0.739	0.3157	73.8%
Filtered Classifier	0.804	0.757	0.760	0.3061	75.7%
LWL	0.834	0.777	0.778	0.2995	77.7%
Decision Table	0.658	0.650	0.653	0.3980	65.0%
DTNB	0.658	0.650	0.653	0.3969	65.0%
NBTree	0.716	0.689	0.693	0.3347	68.9%
Random Forest	0.769	0.767	0.768	0.2483	76.7%

Table 4.5: Results from the external validation test set analysis listed by classifier using all molecular descriptors. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.803	0.650	0.617	0.3426	65.0%
SMO	0.803	0.650	0.617	0.3500	65.0%
J48	0.803	0.650	0.617	0.2776	65.0%
Bagging	0.803	0.650	0.617	0.2953	65.0%
Classification via Regression	0.803	0.650	0.617	0.3047	65.0%
Filtered Classifier	0.803	0.650	0.617	0.2776	65.0%
LWL	0.955	0.950	0.950	0.2510	95.0%
Decision Table	0.803	0.650	0.617	0.4206	65.0%
DTNB	0.803	0.650	0.617	0.4182	65.0%
NBTree	0.803	0.650	0.617	0.2945	65.0%
Random Forest	0.803	0.650	0.617	0.2784	65.0%

Table 4.6: Results from the external validation test set analysis listed by classifier including the molecular descriptors expert selected with cytotoxicity concentration. See Eqs. (1)-(4) for definitions of Precision, Recall, F-Measure, Mean Absolute Error and Accuracy.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.918	0.900	0.900	0.1868	90.0%
SMO	0.803	0.650	0.617	0.3500	65.0%
J48	0.918	0.900	0.900	0.1768	90.0%
Bagging	0.888	0.850	0.849	0.2408	85.0%
Classification via Regression	0.803	0.650	0.617	0.3678	65.0%
Filtered Classifier	0.803	0.650	0.617	0.2776	65.0%
LWL	0.955	0.950	0.950	0.2467	95.0%
Decision Table	0.803	0.650	0.617	0.4206	65.0%
DTNB	0.803	0.650	0.617	0.4182	65.0%
NBTree	0.803	0.650	0.617	0.3082	65.0%
Random Forest	0.888	0.850	0.849	0.2187	85.0%

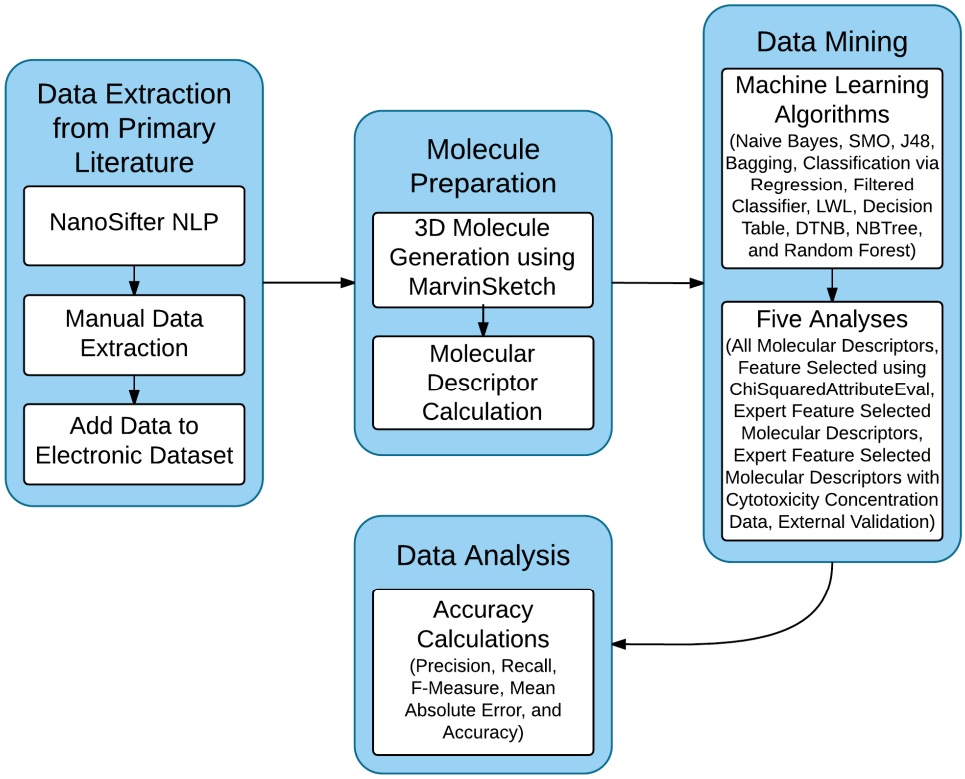

Figure 4.2: Simplified workflow diagram for the method used in this study.

CHAPTER 5

MOLECULAR DYNAMIC SIMULATIONS IN DRUG DELIVERY RESEARCH: CALCIUM CHELATION OF G3.5 PAMAM DENDRIMERS DESCRIPTORS[4]

Abstract

Poly(amido amine) (PAMAM) dendrimers have been considered as possible delivery systems for anticancer drugs. One potential advantage of these carriers would be their use in oral formulations, which will require absorption in the intestinal lumen. This may require the opening of tight junctions by reducing the Ca^{2+} concentration in the intestinal lumen, as has been shown as a possible absorption mechanism in EDTA. Using molecular dynamic simulations, we show that the G3.5 PAMAM dendrimers are able to chelate Ca^{2+} at similar proportions to EDTA, providing support to the hypothesis that oral formulations of PAMAM dendrimers are a plausible approach to deliver highly toxic anticancer agents by an oral route.

Introduction

Poly(amido amine) (PAMAM) dendrimers are complex molecules whose biochemical activity *in vivo* is not fully understood. A particular mechanism of interest is the pathway by which orally taken PAMAM dendrimers may reach their target location

[4] *David E. Jones [a], Albert M. Lund [a, b], Hamidreza Ghandehari [c, d], and Julio C. Facelli [a, d]. [a] Department of Biomedical Informatics, [b] Department of Chemistry, [c] Departments of Bioengineering and Pharmaceutics and Pharmaceutical Chemistry, and [d] Utah Center for Nanomedicine, Nano Institute of Utah, University of Utah, Salt Lake City, UT 84112, US*

when used as nano carriers of anticancer drugs. Specifically, there is interest in understanding how these particles would be able to permeate the tight junctions of the intestinal lumen[147].

Literature consensus shows that tight junctions are dependent upon extracellular calcium (Ca^{2+}) and magnesium for their integrity and function[148]. Extracellular Ca^{2+} is responsible for keeping the tight junctions closed, and it is known that lower concentrations of Ca^{2+} in the intestinal lumen lead to their opening. This has been clearly established using ethylenediaminetetraacetic acid (EDTA), a known Ca^{2+} chelator, that has been shown to open and transverse the tight junctions[149]. Several publications have suggested that PAMAM dendrimers also would be able to travel across the intestinal barrier using the same mechanism[150, 151]. Due to their anionic charge the carboxylic-acid-(COOH) terminated Generation 3.5 (G3.5) PAMAM dendrimers may be capable of chelating Ca^{2+} in solution much like EDTA[148]. Each COOH terminated G3.5 PAMAM dendrimer can theoretically chelate 32 Ca^{2+} ions, thus significantly reducing the extracellular Ca^{2+} and therefore creating an extracellular environment that is prone to opening the tight junctions. This would allow for paracellular transport of G3.5 PAMAM dendrimers via the tight junctions. However, the ability of PAMAM dendrimers to chelate Ca^{2+} has yet to be confirmed by *in vitro* or *in vivo* studies. Simulation studies can be used to better understand this hypothesis and justify more experimental work.

Molecular dynamics (MD) simulations are routinely used to provide understanding and testing of novel hypotheses at the molecular scale for small-molecule drug delivery research[34, 35]. With recent advances in computational power, such simulations can now be used in nanomedicine, for instance to better understand chemical and biological properties of PAMAM dendrimer nanoparticles. Many studies have focused on the use of MD simulations to gain insight on PAMAM dendrimer/ligand conformations and energies. A few molecular docking studies have been reported using MD simulations

to analyze the interactions of siRNA and PAMAM dendrimers[51-53]. Other molecular docking studies have involved PAMAM dendrimers and ligands such as curcumin and porphyrin[54, 55]. These include work by Avila-Salas *et al.*, who used MD simulations and QSAR methods for *in silico* dendrimer-drug affinity studies[56]. Ivanov and Jacobson used MD simulations to test the theoretical possibility of bivalent binding of a dendrimer, covalently appended with multiple copies of a ligand[57]. Lee *et al.* used MD simulations along with chemical analysis to guide the design of a multifunctional PAMAM dendrimer-based nano-therapeutic[58]. Barata *et al.* used MD simulations to identify the key structural design principles for bioactive dendrimer molecules that could be synthesized and biologically evaluated[59]. However, to the authors' knowledge no study has been performed on the ability of Ca^{2+} chelation by PAMAM dendrimers. In this article, we demonstrate how MD simulations can be used to test the ability of G3.5 PAMAM dendrimers to chelate Ca^{2+}. These results have been validated using MD simulations of EDTA as a reference system for Ca^{2+} chelation.

Methods

EDTA and G3.5 PAMAM dendrimer structures were manually constructed using MarvinSketch by ChemAxon[97, 145]. The structures for each molecule were converted to their ionic forms by removing a hydrogen atom from the terminal COOH groups. This resulted in a minus four charge for the EDTA molecule, as there are four terminal COOH groups in it, and a minus 64 charge for the G3.5 PAMAM dendrimer molecule where there are 64 terminal COOH groups.

Force Field and Water Box Preparation

For all simulations, the ff12sb force fields were used along with the general AMBER force field (GAFF) to represent the EDTA and G3.5 PAMAM dendrimer[152]. Four different solvent boxes were prepared for four different simulations: EDTA with Ca^{2+} in water, G3.5

PAMAM dendrimer with Ca^{2+} in water, EDTA with Ca^{2+} in a buffer solution, and G3.5 PAMAM with Ca^{2+} in the same buffer solution. The water model used in all the simulations was the TIP3P[153]. For both simulation studies in water, the concentration of Ca^{2+} was adjusted to approximately 0.115 M, while the simulation studies in the buffer solution were done using concentrations of Ca^{2+} ions of approximately 0.0575 M with the addition of sodium chloride (NaCl) at a concentration of approximately 0.115 M. The concentrations used in the buffer simulations are representative of the experimental solutions used to examine the ability of EDTA to chelate Ca^{2+} and increase epithelial absorption by Vllasaliu *et al.* and Tomita *et al.*[154, 155]. The simulations in the buffer solution have been performed to further test the ability of Ca^{2+} chelation by EDTA and G3.5 PAMAM dendrimers when a competing ion (Na^{+}) is present in the solution. For the EDTA simulations the volume of the water box was approximately 75,000 $Å^3$, while for the simulations with G3.5 PAMAM dendrimer the volume of the water box was approximately 460,000 $Å^3$. All the simulations were performed using periodical binding conditions and a nonbonded cutoff of 9.0 Å.

Three independent MD simulations were performed for each system to obtain ensemble averages of multiple MD runs. Before each simulation all of the counter ions were randomly distributed in the water box at 8.00 Å from the molecule of interest, EDTA or G3.5 PAMAM dendrimer, using cpptraj[156]. The next three subsections describe the initial minimization, 40 ps equilibration, and 30 ns MD simulation parameters used in all MD simulations.

Initial Minimization

The initial minimization consisted of a total of 3,000 minimization iterations, of which 1,000 were done using the steepest descent method and 2,000 using conjugate gradients.

40 ps Equilibration

Following the minimization, the 40 ps equilibration phase consisted of 20,000 steps with a time step of 2 fs. For the equilibration period the temperature was increased from 0.0 to 300.0 K. and Langevin dynamics was used with a collision frequency of 1.0 ps^{-1}.

30 ns MD Simulation

Following equilibration, the 30 ns MD simulation consisted of 1.5 x 10^7 steps with a time step of 2 fs. For the MD simulation the temperature and pressure were held constant at 300.0 K and 1.0 bar, respectively, and Langevin dynamics was used with a collision frequency of 1.0 ps^{-1}.

Analysis

Trajectory analysis of the 30 ns MD simulation was performed using cpptraj [156]. The water molecules were removed from the trajectories and a custom Python script was used to plot the distance from each counter ion to the Van der Walls surface of the molecule of interest (EDTA or G3.5 PAMAM dendrimer) for each time step of the 30 ns MD simulation. This script also calculated the average distance and minimum distance of each counter ion to the surface of the molecule of interest (EDTA or G3.5 PAMAM dendrimer) during the 30 ns MD simulation and the percentage of dwell time, that is the number of steps in which an individual counter ion is within 3.0 Å of the surface of the molecule divided by the total number of steps in the MD simulation.

Apt Computational Environment

All calculations were performed on the Tangent cluster at the Center for High Performance Computing. Tangent is a part of the Adaptable Profile-Driven Testbed (Apt), an initiative by the Flux Research Group to provide a flexible, on-demand cloud computing

environment targeted at researchers and scientists (https://www.flux.utah.edu/project/apt).

We used our research project as a demonstration of how to use a computationally intensive application in a cloud computing environment. This study shows that it is possible to perform computationally significant drug delivery research in such an environment, which may provide researchers an appealing and more cost-effective alternative to traditional, dedicated high performance computing (HPC) environments.

Results and Discussion

As discussed above, four different simulations were carried out: EDTA and Ca^{2+} in water, G3.5 PAMAM dendrimer and Ca^{2+} in water, EDTA and Ca^{2+} in a buffer, and G3.5 PAMAM dendrimer and Ca^{2+} in the same buffer. The results from each of these simulations as well as the performance of the Apt computational environment are described in the following subsections. In all cases our results are reported as an average of the values obtained in each of the three independent MD simulations performed for each system to represent ensemble average values.

MD Simulation Study of EDTA and Ca^{2+} in Water

Results from the EDTA and Ca^{2+} in water MD simulations are presented in Figure 5.1 and Table 5.1. Figure 5.1 presents the three-dimensional depiction of the final recorded step of one of the MD simulations of EDTA and Ca^{2+} in water, where it is apparent that the Ca^{2+} atoms are attached to the EDTA molecule. Table 5.1 shows the values of the averages of the distance, minimum distance, and percentage dwell time of the counter ions (Cl^- and Ca^{2+}) included in the simulations. The average values and their standard deviations are the average over all the ions of the same type over all the trajectories from the three MD simulations performed here for this system.

Both the figure and table present results that are indicative of the ability of EDTA

to chelate Ca^{2+} in water. Figure 5.1 presents a clear depiction of two Ca^{2+} ions interacting with the EDTA molecule in the final recorded step of one of the MD simulation runs. As can be seen in Table 5.1, the average minimum distance, 2.86 Å, from the surface of EDTA for the Ca^{2+} ions shows that in all the simulations the Ca^{2+} ions have migrated, at least temporarily, from a distance of 8.00 Å to a distance within 3.00 Å of the surface of the EDTA molecule. The average distance of the Ca^{2+} ions to the surface of the EDTA molecule is comparably large, but this is largely due to the fact that EDTA has only four terminal COOH groups. This allows for only two Ca^{2+} ions to bind to EDTA. To ensure that the same concentration of Ca^{2+} ions was used for all the MD simulation studies in water, 5 Ca^{2+} ions were used in the EDTA and Ca^{2+} in water MD simulations, therefore not all the Ca^{2+} ions can be in close proximity to the EDTA molecule. This resulted in a larger average distance and standard deviation. The average value of approximately 10 Å is consistent with a situation in which, on average, three Ca^{2+} ions are at an average distance of ~15 Å, which is the average distance for the Cl ions. The other two Ca^{2+} ions are at an average distance of ~2.5 Å, which corresponds to a situation in which the ions are bound to EDTA. This is confirmed by the average percentage dwell time for Ca^{2+} ions, which shows that for most of the simulation, 80% of the time, at least two of the Ca^{2+} ions are interacting with the surface of EDTA. These simulations' results agree with the experimental results showing the large affinity of EDTA to chelate Ca^{2+}[149].

The small standard deviations observed for the average minimum distance indicate that all ions at a given time in the simulation visit the proximity of the EDTA molecule, while the large standard deviations observed both for the average distance and average percentage dwell time of Ca^{2+} are an indication that there is dynamic equilibrium in which preferentially these ions are close to the EDTA surface. The small deviation on the Cl^- ion values indicates that these ions remain far from the EDTA surface, an expected result based on the EDTA and Cl^- polarity.

MD Simulation Study of EDTA and Ca^{2+} in a Buffer Solution

Results from the MD simulation study of EDTA and Ca^{2+} in a buffer are presented in Table 5.2, which shows the average distance, minimum distance, and average percentage dwell time of all the counter ions (Cl^-, Na^+, and Ca^{2+}) in the buffer solution, as well as their standard deviations of these values when averaged over all the trajectories in the three simulations performed here for this system.

The average minimum distance of Ca^{2+} ions to the surface of the EDTA molecule is 2.54 Å, slightly smaller than the one observed in the water simulation and within 3.00 Å of the surface of the EDTA molecule. The average distance of the Ca^{2+} ions is 9.05 Å, also slightly smaller than in the water simulations. This relatively large value can be explained by the same argument used for the simulations in water. The average dwell time for the Ca^{2+} ions also indicates that on average two Ca^{2+} ions are interacting with the surface of EDTA for a majority of the MD simulations. The results in the table show that the Na^+ ions are much less likely to be close to the EDTA surface. It is not surprising that the minimum distance of Na^+ (2.02 Å) is closer than that observed in the Ca^{2+}, since the Na^+ ions are smaller and they only interact with one surface group on the EDTA molecule, whereas the Ca^{2+} ions commonly interact with two surface groups on the EDTA molecule. The average distance and the average percentage dwell time are considerably larger and smaller, respectively. These results show that while the Na^+ is able to get in proximity to the EDTA, its interaction is less favorable than for Ca^{+2}, as demonstrated by the larger average distance and smaller percentage dwell time.

The results in the buffer solution are similar to those observed in the simulations of EDTA and Ca^{2+} in water. While it can be observed in the simulations that there is some competition for binding sites on EDTA between the 6 Na^+ ions and 2 Ca^{2+} ions in the simulations, when comparing the results in Tables 5.1 and 5.2 it is apparent that this competition did not affect EDTA's ability to chelate Ca^{2+}. This indicates that the ability of

EDTA to chelate Ca^{2+} is not affected by the presence of a competing ion (Na^{+}) in the simulation environment. The results of the MD simulations are consistent with experimental results discussed above[154, 155].

The simulations results agree with the experimental results showing the ability of EDTA to chelate Ca^{2+}[149] and validate our modeling approach to study chelation of G3.5 PAMAM dendrimers.

MD Simulation Study of G3.5 PAMAM Dendrimer and Ca^{2+} in Water

Results from the MD simulations of G3.5 PAMAM dendrimer and Ca^{2+} are presented in Figure 5.2 and Table 5.3. Figure 5.2 shows a three-dimensional depiction of the final recorded step of one of the simulations of G3.5 PAMAM dendrimer with Ca^{2+} in water, while Table 5.3 shows the values for the average distance and minimum distance from the surface of the G3.5 PAMAM dendrimer molecule and average percentage dwell time of the Ca^{2+} ions, as well as their standard deviations calculated over all the trajectories in the three independent simulations performed for this system.

Both the figure and table present results that clearly show the ability of G3.5 PAMAM dendrimer to chelate Ca^{2+}. Figure 5.2 shows that multiple Ca^{2+} ions are interacting with the surface of the G3.5 PAMAM dendrimer molecule in the final recorded step of one of the MD simulation runs. As can be seen in Table 5.3, the average minimum distance of Ca^{2+} from the surface of G3.5 PAMAM dendrimer is 2.40 Å, which is closer than the average minimum distance of Ca^{2+} observed in the EDTA simulations in water (2.86 Å). Also, the average distance of the Ca^{2+} ions to the surface of the G3.5 PAMAM dendrimer molecule, 4.13 Å, is much smaller than that observed with EDTA (10.34 Å) in water, consistent with the fact that in this system all Ca^{2+} ions can be bound to the G3.5 PAMAM at the same time. The average percentage dwell time of 0.86 shows that the G3.5 PAMAM dendrimer is nearly always, almost ninety percent of the time, binding Ca^{2+} ions.

Note that the decrease in the average distance and dwell time is also associated with the fact that the calcium ions in this simulation are in the correct stoichiometric ratio, such that all the ions can be binding to the G3.5 PAMAM dendrimer at the same time.

MD Simulation Study of G3.5 PAMAM Dendrimer and Ca^{2+} in a Buffer

Results from the MD simulation study of the G3.5 PAMAM dendrimer and Ca^{2+} in a buffer are presented in Figure 5.3 and Table 5.4. Figure 5.3 shows the plots of the distance of the counter ions (Ca^{2+}) from the surface of the G3.5 PAMAM dendrimer molecule versus time during three independent MD runs. Table 5.4 shows the values for the average distance and minimum distance from the surface of the G3.5 PAMAM dendrimer molecule and the percentage dwell time of all the counter ions (Na^{+} and Ca^{2+}) present in the buffer solution, as well as their standard deviations calculated over all trajectories from the three independent MD simulations performed in this system.

Figure 5.3 shows that for the three independent MD runs performed, at approximately halfway through the MD simulations all 16 Ca^{2+} ions are bound to the surface of the G3.5 PAMAM dendrimer. Table 5.4 shows results almost the same as those obtained in the water simulations and demonstrates preferential binding of G3.5 PAMAM dendrimer to Ca^{2+} ions. These results show that the results obtained in water can also be expected in a buffer solution that mimics a situation closer to the intestinal lumen milieu, and that the preferential binding is not sensitive to the details of the solution used in the simulations. Moreover, the preferential binding of Ca^{+2} is highlighted by the difference between the average distance of each ion to the surface of the G3.5 PAMAM. The average distance of the Ca^{2+} ions to the surface of the G3.5 PAMAM dendrimer molecule, 3.55 Å, is much smaller than that observed with the Na^{+} ions, 7.50 Å. This indicates that the Ca^{2+} ions appear to be more frequently bound to the G3.5 PAMAM dendrimer than the Na^{+} ions. The average minimum distance of Ca^{2+} from the surface of G3.5 PAMAM dendrimer

of 2.40 Å is consistent with what was observed in the G3.5 PAMAM dendrimer and Ca^{2+} simulation in water. The average minimum distance of Na^{+} (2.02 Å) is shorter than that observed in the Ca^{2+}, in agreement with the results discussed above for the simulation of EDTA in buffer. Results of the average percentage dwell time of the Na^{+} and Ca^{2+} ions show a relatively large difference, 0.60 for Na^{+} and 0.91 for Ca^{2+} ions, which is similar to that observed in the simulation in water. This indicates that the Ca^{2+} ions appear to be more frequently bound to the G3.5 PAMAM dendrimer than the Na^{+} ions. Results show that Ca^{2+} is an effective chelation agent for G3.5 PAMAM dendrimer, indicating that under the simulation conditions the G3.5 dendrimer is a Ca^{2+} chelator in both water and in a buffer solution.

Apt Computational Environment

The use of AMBER in the Apt HPC environment has been extremely successful. During the course of these studies we were able to secure all the needed resources for the project, and because Apt is able to mirror a standard HPC environment, the migration to this cloud-based computing environment has been straightforward. Moreover, the overhead for using the Apt environment has been minimum. The only overhead incurred when using Apt is the instantiation and de-provisioning of the infrastructure, which is typically less than 15 minutes, a small fraction of a typical AMBER simulation.

Conclusion

Using MD simulations validated by agreement with existing experimental results in EDTA, we have shown that in MD simulations G3.5 PAMAM dendrimers are capable of chelating Ca^{2+} and therefore they may be good candidates for oral formulations that require opening the tight junctions for absorption. We have demonstrated that computationally intensive applications of interest in drug delivery research can effectively use cloud computing environments like Apt.

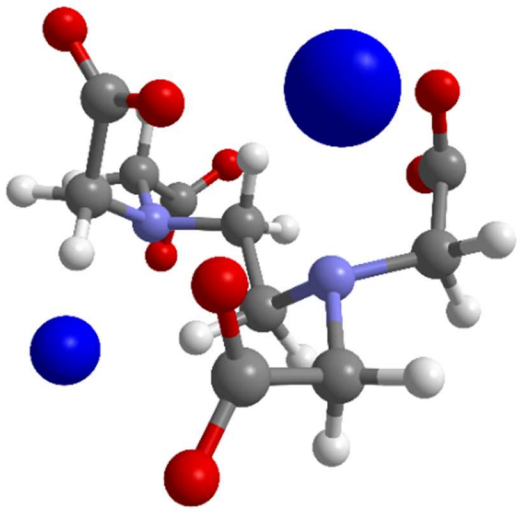

Figure 5.1: Three-dimensional representation of the final recorded step of one of the MD simulations of the EDTA and Ca^{2+} in water. The blue spheres represent Ca^{2+} ions in close proximity to the EDTA molecule.

Table 5.1: Average distance and average minimum distance from the van der Walls surface of the EDTA molecule and average percentage dwell time of the counter ions (Cl^- and Ca^{2+}) included in this simulation. Standard deviations, over the three independent runs of the EDTA and Ca^{2+} in water MD simulations, given between brackets.

Counter Ion	**Average Distance (Å)**	**Average Minimum Distance (Å)**	**Average % Dwell Time**
Cl^-	15.89 (0.47)	2.95 (0.21)	0.00 (0.0)
Ca^{2+}	10.34 (5.70)	2.86 (0.58)	0.39 (0.42)

Table 5.2: Average distance and average minimum distance from the van der Walls surface of the EDTA molecule and average percentage dwell time of the counter ions (Cl^-, Na^+ and Ca^{2+}) included in this simulation. Standard deviations, over the three independent runs of the EDTA and Ca^{2+} in a buffer MD simulation, given between brackets.

Counter Ion	**Average Distance (Å)**	**Average Minimum Distance (Å)**	**Average % Dwell Time**
Cl^-	15.84 (0.43)	3.05 (0.26)	0.00 (0.0)
Na^+	12.26 (4.04)	2.02 (0.09)	0.24 (0.29)
Ca^{2+}	9.05 (4.96)	2.54 (0.40)	0.47 (0.37)

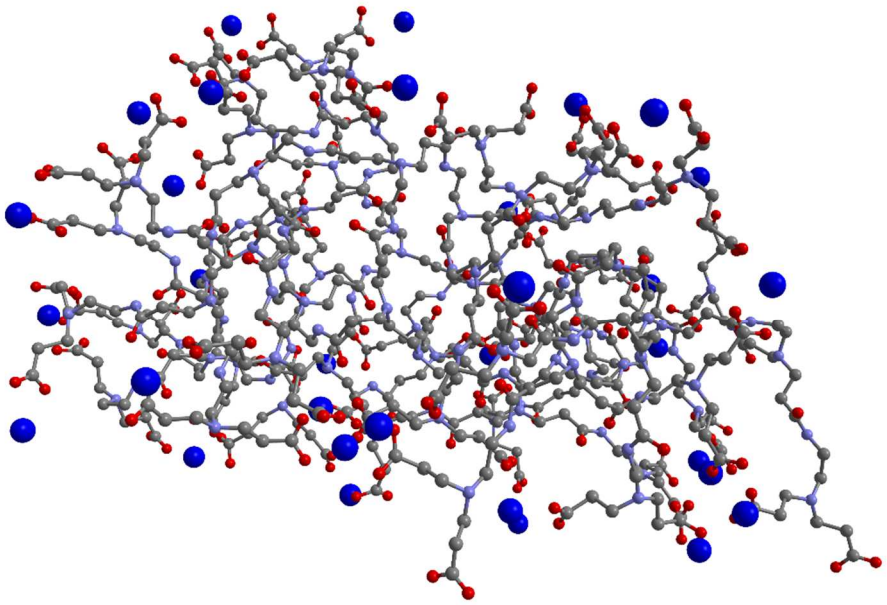

Figure 5.2: Three-dimensional representation of the final recorded step of one of the MD simulations of the G3.5 PAMAM dendrimer and Ca^{2+} in water. The blue spheres represent Ca^{2+} ions present in close proximity to the G3.5 PAMAM dendrimer.

Table 5.3: Average distance and average minimum distance from the van der Walls surface of the G3.5 PAMAM molecule and average percentage dwell time of the counter ion (Ca^{2+}) included in this simulation. Standard deviations, over the three independent runs of the G3.5 PAMAM dendrimer and Ca^{2+} in water MD simulation, given between brackets.

Counter Ion	**Average Distance (Å)**	**Average Minimum Distance (Å)**	**Average % Dwell Time**
Ca^{2+}	4.13 (2.83)	2.40 (0.02)	0.86 (0.22)

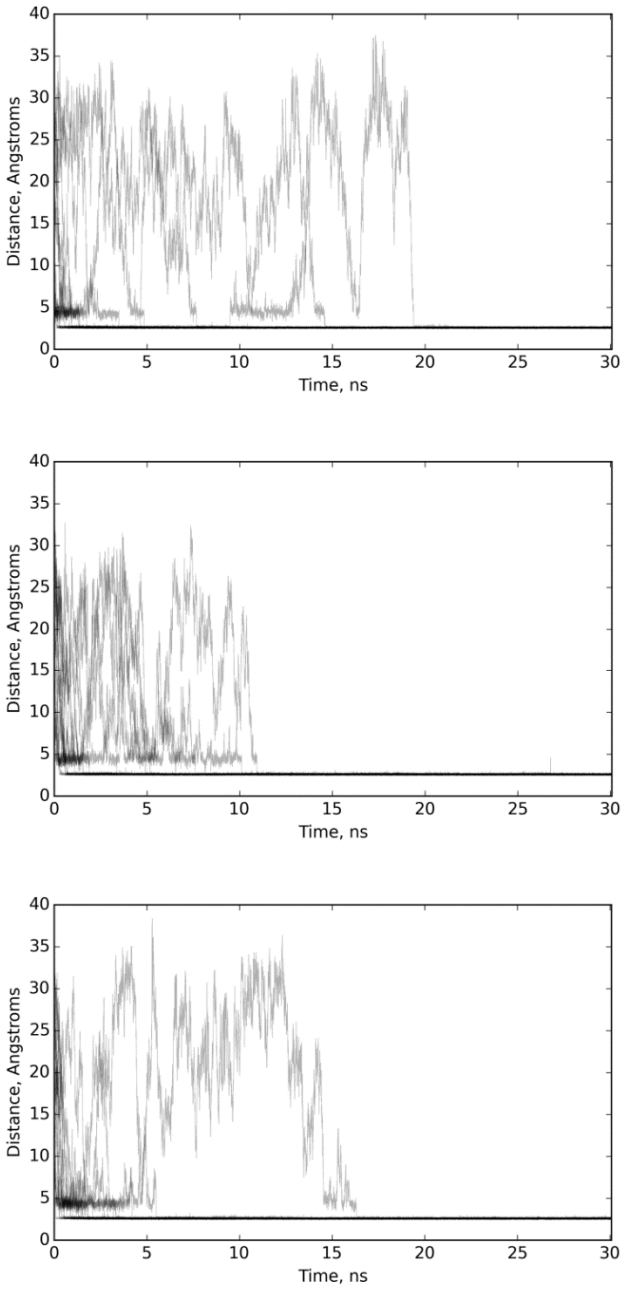

Figure 5.3: Plots of the distance of the counter ions (16 Ca^{2+} ions) from the van der Walls surface of the G3.5 PAMAM dendrimer molecule versus time for the three independent runs of the MD simulations of the G3.5 PAMAM dendrimer and Ca^{2+} in a buffer solution.

Table 5.4: Average distance and average minimum distance from the van der Walls surface of the G3.5 PAMAM molecule and average percentage dwell time of the counter ion (Na^+ and Ca^{2+}) included in this simulation. Standard deviations, over the three independent runs of the G3.5 PAMAM dendrimer and Ca^{2+} in a buffer MD simulation, given between brackets.

Counter Ion	**Average Distance (Å)**	**Average Minimum Distance (Å)**	**Average % Dwell Time**
Na^+	7.50 (3.13)	2.02 (0.05)	0.60 (0.20)
Ca^{2+}	3.55 (2.20)	2.40 (0.01)	0.91 (0.13)

CHAPTER 6

CONCLUSIONS

Importance of Research Findings

The development of the field of nanoinformatics is a direct result of the steady growth of the field of nanomedicine and nanotechnology. Nanoinformatics is a field that consists of several areas of research that use different informatics methodologies. This dissertation focuses on the use of three different informatics methodologies to assist in the development of PAMAM dendrimer nanoparticle drug delivery systems: information extraction (IE), specifically natural language processing (NLP); data mining and machine learning; and molecular dynamics (MD) simulation.

A literature review was conducted to examine the strides that have been made using data mining and machine learning to develop nano-QSARs and other methods to predict both functional and structural properties of nanoparticles in the field of nanomedicine. Journal articles have been published attempting to predict several nanoparticle properties, including cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanoparticle size, and polydispersity. The most common use of data mining and machine learning was to explore cytotoxicity caused by inorganic nanoparticles, and it was found that the most common factors determining cytotoxicity are charge, concentration, and size; this is not surprising as these properties have been hypothesized to be important indications of the potential cytotoxicity of nanoparticles[98]. Two areas lacking adequate research became apparent: the use of data mining and machine learning methods to predict cytotoxicity of organic nanoparticles and

to analyze *in vivo* data regarding nanoparticles.

A method, NanoSifter, based upon NLP methodologies, was developed to automatically extract numeric values associated with dendrimer property terms from the nanomedicine literature. NanoSifter's results illustrate that this NLP system is both reliable and accurate at extracting information regarding dendrimers from the cancer nanomedicine literature. The results from this system are promising for the future of IE in the field of nanoinformatics.

Next, data mining and machine learning were explored for their use in predicting cytotoxicity in Caco-2 cells caused by poly(amido amine) (PAMAM) dendrimers. As was observed in many of the research articles in the review, the results of the classification method used show good accuracy. Also, it was observed that properties regarding charge, size, and concentration of the PAMAM dendrimers were the most important factors in the classification of cytotoxicity in Caco-2 cells treated with PAMAM dendrimers.

Finally, the use of MD simulations was examined for the utility of determining whether or not generation 3.5 (G3.5) PAMAM dendrimers can chelate calcium (Ca^{2+}). In order to validate the MD simulation model, EDTA was used as a gold standard molecule because it has been experimentally shown to chelate Ca^{2+}. Results from the MD simulations show that G3.5 PAMAM dendrimers are capable of chelating Ca^{2+} and therefore they may be good candidates for oral formulations that require opening the tight junctions for absorption.

Contribution to the Field

The chapters and research presented in this dissertation clearly illustrate the power and accuracy that nanoinformatics methods and techniques can bring toward gaining further knowledge and understanding in the field of nanomedicine. This dissertation examined the use of three very different informatics methodologies: NLP,

data mining and machine learning, and MD simulation. Research presented in this dissertation serves as a proof of concept for applying nanoinformatics methods and techniques to gain further knowledge and understanding of PAMAM dendrimers, which was successfully accomplished by the methodologies explored in this dissertation both individually and in combination. Therefore, these methods and techniques should in theory be capable of expansion to a variety of different nanoparticles used for applications in nanomedicine. The ability to reliably predict and simulate PK/PD properties of orally delivered PAMAM dendrimer nanoparticle drug delivery systems using *in silico* approaches has the potential for high payoff in nanomaterial development, allowing the concentration of scarce development resources into the synthesis and confirmatory testing of promising PAMAM dendrimers.

Future Research

The potential for future research in the field of nanoinformatics is vast. Since it is a relatively young field, there are still many areas within nanoinformatics that have yet to be explored. There are several directions that could be taken for further advancing the research presented in this dissertation.

With regard to the NanoSifter NLP system, there are a few directions of future research that could be taken. The obvious, low-hanging fruit would be to improve upon the system in order to predominantly improve our precision and f-measure values, while maintaining our recall. To accomplish this, it is necessary to develop a text classification method to further enhance NanoSifter. Two specific foci of this text classification method would be to separate identification of chemical entities and nanoparticles so that information extracted from the literature is specific to the nanoparticle of interest, and further analysis of percentages within the text to improve the precision achieved for the properties of cell viability and transfection efficiency. Two other obvious goals would be to

expand this IE method to include more properties and nanoparticle subclasses to be annotated and extracted from the literature. Something that would be very powerful would be to more seamlessly integrate the NPO into our system so that the annotations and extractions contain descriptive metadata. Another goal is the implementation of a negation analysis tool into the NanoSifter NLP system, because this could help capture data regarding instances where an article states that the dendrimer nanoparticles were not toxic at a certain concentration. Finally, developing and implementing a method to annotate and extract information from figures and tables would be very useful, as many times valuable information within the nanomedicine literature is contained in figures and tables rather than the body of text.

Both the review and the research article regarding the use of data mining and machine learning to predict cytotoxicity of PAMAM dendrimers in Caco-2 cells indicate several potential areas for further research expansion. First of all, the methods presented in the research article could be expanded to analyze and predict other biochemically relevant properties of not only unmodified PAMAM dendrimers but also surface-modified PAMAM dendrimers. Also, it is necessary to improve the reliability and robustness of our classification model by conducting further studies, including the collection of much larger datasets. As with the IE methods, an obvious goal is to explore utilizing data mining and machine learning to predict many PK/PD properties of a variety of nanoparticle subclasses. A specific focus would be to look at using data mining and machine learning to predict PK/PD properties *in vivo*, because there is a lack of research in this area. Another future study which would improve upon the research article regarding the use of data mining and machine learning to predict cytotoxicity of PAMAM dendrimers in Caco-2 cells would be to utilize the same dataset and perform an unsupervised machine learning (clustering) method to statistically validate which properties of the 51 calculated molecular descriptors are truly indicative of cytotoxicity.

Some very interesting studies could be carried out utilizing MD and coarse-grained simulations. With regard to PAMAM dendrimers, using simulation methods to look at how these molecules interact with the cells of the intestinal epithelium and potentially pass through the tight junctions is currently being explored by colleagues. Another interesting simulation study would be to use simulation methods to analyze how a corona of ions and molecules build up around PAMAM dendrimers and how it effects PK/PD properties of the molecule. A final potential simulation study would be to expand the use of simulation methods to research and assess a variety of PK/PD properties of other nanoparticle subclasses.

Another area of future work could be to focus on further development of standards within the field of nanoinformatics and nanomedicine. The information and knowledge gained from this nanoinformatics research, as well as others' work, should and could be compiled for well-established subclasses of nanoparticles to assist in the development of standards. Standardization is vital for the advancement of both nanoinformatics and nanomedicine.

As stated previously, there will be challenges along the way that we as a nanoinformatics community will need to overcome. Currently some of the most pressing challenges are the lack of high-quality experimental data, lack of knowledge regarding interactions between nanoparticles, such as aggregation and high polydispersity in nanoparticles[121, 122]. It is my belief that in working together with the nanomedicine community, we will be able to combat these issues and propel both the fields of nanoinformatics and nanomedicine.

APPENDIX

Table A1: Listing of the molecular descriptors and their definitions selected from MarvinSketch[97].

Molecular Descriptor	**Units**	**Definition**
Molecular Weight	Da	Average molecular mass calculated from the standard atomic weights [157].
Exact Mass	Da	Monoisotopic mass calculated from the weights of the most abundant natural isotopes of the elements [158].
Atom Count		Number of all atoms in the molecule.
pI		Net charge of an ionizable molecule is zero at a certain pH. This pH is called the isoelectric point, also referred to as pI.
logP		The octanol/water partition coefficient, which is used in quantitative structure activity relationships (QSAR) analysis and rational drug design as a measure of molecular hydrophobicity [159].
logD		The octanol-water distribution coefficient, log*D* represents the compounds at any pH value.
Molecular Polarizability	$Å^3$	The electric field generated by partial charges of a molecule spread through intermolecular cavities and the solvent. The induced partial charge (induced dipole) has a tendency to diminish the external electric field. This phenomenon is called polarizability.
Aliphatic Atom Count		Number of atoms in the molecule having no aromatic bond (excluding hydrogens).
Aliphatic Bond Count		Number of non-aromatic bonds in the molecule (excluding bonds of hydrogen atoms).
Aromatic Atom Count		Number of atoms in the molecule having aromatic bonds.
Aromatic Bond Count		Number of aromatic bonds in the molecule.
Asymetric Atom Count		The number of asymmetric atoms (having four different ligands).
Bond Count		Number of bonds in the molecule including bonds of hydrogen atoms.
Chain Atom Count		Number of chain atoms (non-ring atoms excluding hydrogens).
Chain Bond Count		Number of chain bonds (non-ring bonds excluding bonds of hydrogen atoms).
Chiral Center Count		The number of tetrahedral stereogenic centers. This function identifies two chiral centers in 1,4-dimethylcyclohexane, which does not contain asymmetric atoms.
Ring Atom Count		Number of ring atoms.
Ring Bond Count		Number of ring bonds.
Rotatable Bond Count		Number of rotatable bonds in the molecule. Unsaturated bonds, and single bonds connected to hydrogens or terminal atoms, single bonds of amides, sulphonamides and those connecting two hindered aromatic rings (having at least three ortho substituents) are considered non-rotatable.
Stereo Double Bond Count		Number of double bonds with defined stereochemistry.
Aliphatic Ring Count		Number of those rings in the molecule that have non-aromatic bonds (SSSR based).

Table A1: Continued

Molecular Descriptor	Units	Definition
Aromatic Ring Count		Number of aromatic rings in the molecule. This number is calculated from the smallest set of smallest aromatic rings (SSSAR), which might contain rings which are not part of the standard SSSR ring set. As a consequence, the sum of the aliphatic ring count and the aromatic ring count can sometimes be greater than the ring count value. The difference is the signal of a macroaromatic ring system.
Carbo Ring Count		Number of rings containing only carbon atoms.
Carboaliphatic Ring Count		Number of aliphatic rings containing only carbon atoms.
Carboaromatic Ring Count		Number of aromatic rings containing only carbon atoms (SSSAR based).
Fused Aliphatic Ring Count		Number of aliphatic rings having common bonds with other rings.
Fused Aromatic Ring Count		Number of aromatic rings having common bonds with other rings.
Fused Ring Count		Number of fused rings in the molecule (having common bonds).
Hetero Ring Count		Number of rings containing hetero atom(s).
Heteroaliphatic Ring Count		Number of aliphatic heterocycles in the molecule.
Heteroaromatic Ring Count		Number of aromatic heterocycles in the molecule.
Largest Ring Size		Size of the largest ring in the molecule.
Largest Ring System Size		Number of rings in the largest ring system.
Ring Count		Number of rings in the molecule. This calculation is based on SSSR (Smallest Set of Smallest Rings).
Ring System Count		Number of disjunct ring systems.
Smallest Ring Size		Size of the smallest ring in the molecule.
Smallest Ring System Size		Number of rings in the smallest ring system.
Platt Index		Sum of the edge degrees of a molecular graph.
Randic Index		Harmonic sum of the geometric means of the node degrees for each edge.
Harary Index		Half-sum of the off-diagonal elements of the reciprocal molecular distance matrix of the molecule.
Hyper Wiener Index		A variant of the Wiener index.
Szeged Index		The Szeged index extends the Wiener index for cyclic graphs by counting the number of atoms on both sides of each bond (those atoms only which are nearer to the given side of the bond than to the other), and sum these counts.
Wiener Index		The average topological atom distance (half of the sum of all atom distances) in the molecule.
Wiener Polarity		The number of 3 bond length distances in the molecule.
Cyclomatic Number		The smallest number of bonds which must be removed so that no circuit remains. Also known as circuit rank.
Fragment Count		Number of fragments in the sketch.
H-Bond Donor Count		Hydrogen Bond Donor calculates atomic hydrogen bond donor inclination.
H-Bond Donor Sites		Hydrogen Bond Donor calculates atomic hydrogen bond donor inclination.
H-Bond Acceptor Count		Hydrogen Bond Acceptor calculates atomic hydrogen bond acceptor inclination.
H-Bond Acceptor Sites		Hydrogen Bond Acceptor calculates atomic hydrogen bond acceptor inclination.
Refractivity	$10^6[m^3mol^{-1}]$	Molar refractivity is strongly related to the volume of the molecules and to London dispersive forces that has important effect in drug-receptor interaction.

Table A2: Results from the leave-one-out cross-validation listed by classifier of the first analysis including all molecular descriptors.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.738	0.738	0.725	0.2972	73.8%
SMO	0.780	0.767	0.751	0.2330	76.7%
J48	0.748	0.718	0.722	0.3180	71.8%
Bagging	0.800	0.777	0.780	0.3241	77.7%
Classification via Regression	0.780	0.767	0.751	0.2956	76.7%
Filtered Classifier	0.748	0.718	0.722	0.3180	71.8%
LWL	0.834	0.777	0.778	0.2971	77.7%
Decision Table	0.698	0.680	0.683	0.3746	68.0%
DTNB	0.755	0.728	0.732	0.3145	72.8%
NBTree	0.834	0.777	0.778	0.2520	77.7%
Random Forest	0.750	0.738	0.741	0.2874	73.8%

Table A3: Results from the leave-one-out cross-validation listed by classifier of the second analysis including the automatically feature selected molecular descriptors.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.738	0.738	0.725	0.2972	73.8%
SMO	0.780	0.767	0.751	0.2330	76.7%
J48	0.748	0.718	0.722	0.3180	71.8%
Bagging	0.755	0.728	0.732	0.3241	72.8%
Classification via Regression	0.780	0.767	0.751	0.2956	76.7%
Filtered Classifier	0.748	0.718	0.722	0.3180	71.8%
LWL	0.834	0.777	0.778	0.2971	77.7%
Decision Table	0.650	0.641	0.644	0.3736	64.1%
DTNB	0.755	0.728	0.732	0.3145	72.8%
NBTree	0.834	0.777	0.778	0.2520	77.7%
Random Forest	0.750	0.738	0.741	0.2879	73.8%

Table A4: Results from the leave-one-out cross-validation listed by classifier for the third analysis including the molecular descriptors selected by experts.

Classifier	Precision	Recall	F-Measure	Mean Absolute Error	Accuracy
Naïve Bayes	0.796	0.786	0.789	0.2550	78.6%
SMO	0.738	0.738	0.725	0.2621	73.8%
J48	0.748	0.718	0.722	0.3180	71.8%
Bagging	0.813	0.786	0.789	0.3208	78.6%
Classification via Regression	0.796	0.786	0.789	0.2954	78.6%
Filtered Classifier	0.834	0.777	0.778	0.3051	77.7%
LWL	0.834	0.777	0.778	0.3030	77.7%
Decision Table	0.604	0.602	0.603	0.3867	60.2%
DTNB	0.604	0.602	0.603	0.3829	60.2%
NBTree	0.834	0.777	0.778	0.3271	77.7%
Random Forest	0.750	0.738	0.741	0.2865	73.8%

Table A5: J48 classification accuracy and RMS when using the features selected using all possible WEKA recommended pairs of Attribute Evaluator and Search Method. The selected features are given in the third column.

Attribute Evaluator	Search Method	Selected Attributes	J48 Accuracy	J48 RMS Error
CfsSubsetEval	BestFirst	pI, logP	74.8	0.4163
CfsSubsetEval	ExhaustiveSearch	DID NOT RUN DUE TO THE LARGE SIZE OF FEATUE SET		
CfsSubsetEval	GreedyStepwise	pI, logP	74.8	0.4163
ChiSquaredAtributeEval	Ranker	H-Bond_Acceptor_Sites, pI, logP, Harary_Index, Refractivity, Bond_Count, Molecular_Polarizability, Rotatable_Bond_Count, Atom_Count, logD, Aliphatic_Bond_Count, Chain_Bond_Count, Chain_Atom_Count, Aliphatic_Atom_Count, Exact_Mass, Molecular_Weight, Wiener_Index, Randic_Index, Szeged_Index, Wiener_Polarity, Platt_Index, H-Bond_Donor_Count, Hyper_Wiener_Index, H-Bond_Donor_Sites, H-Bond_Acceptor_Count	74.8	0.401
ClassifierSubsetEval	GreedyStepwise	NO ATTRIBUTES SELECTED		
ConsistencySubsetEval	GreedyStepwise	pI, logD	74.8	0.4163
CostSensitiveAttributeEval	Ranker	NO ATTRIBUTES SELECTED		
CostSensitiveSubsetEval	GreedyStepwise	NO ATTRIBUTES SELECTED		
FilteredAttributeEval	Ranker	H-Bond_Acceptor_Sites, Harary_Index, logP, pI, Bond_Count, Refractivity, Molecular_Polarizability, Rotatable_Bond_Count, Atom_Count, logD, H-Bond_Donor_Sites, Aliphatic_Atom_Count, Chain_Bond_Count, Aliphatic_Bond_Count, Exact_Mass, Chain_Atom_Count, Molecular_Weight, Szeged_Index, Wiener_Polarity, Randic_Index, Wiener_Index, Platt_Index, H-Bond_Donor_Count, Hyper_Wiener_Index, H-Bond_Acceptor_Count	74.8	0.401
FilteredSubsetEval	GreedyStepwise	pI, logP	74.8	0.4163
GainRatioAttributeEval	Ranker	pI, Hyper_Wiener_Index, logP, Platt_Index, Aliphatic_Atom_Count, Szeged_Index, Aliphatic_Bond_Count, Chain_Atom_Count, Randic_Index, Molecular_Weight, Chain_Bond_Count, Wiener_Polarity, H-Bond_Donor_Count, Exact_Mass, Wiener_Index, logD, Harary_Index, Bond_Count, Rotatable_Bond_Count, Atom_Count, Molecular_Polarizability, Refractivity, H-Bond_Acceptor_Sites, H-Bond_Donor_Sites, H-Bond_Acceptor_Count	74.8	0.401

Table A5: Continued

Attribute Evaluator	Search Method	Selected Attributes	J48 Accuracy	J48 RMS Error
InfoGainAttributeEval	Ranker	H-Bond_Acceptor_Sites, Harary_Index, logP, pI, Bond_Count, Refractivity, Molecular_Polarizability, Rotatable_Bond_Count, Atom_Count, logD, H-Bond_Donor_Sites, Aliphatic_Atom_Count, Chain_Bond_Count, Aliphatic_Bond_Count, Exact_Mass, Chain_Atom_Count, Molecular_Weight, Szeged_Index, Wiener_Polarity, Randic_Index, Wiener_Index, Platt_Index, H-Bond_Donor_Count, Hyper_Wiener_Index, H-Bond_Acceptor_Count	74.8	0.401
LatentSemanticAnalysis	Ranker	Molecular_Weight, Exact_Mass, Atom_Count, pI, logP, logD, Molecular_Polarizability, Aliphatic_Atom_Count, Aliphatic_Bond_Count, Bond_Count, Chain_Atom_Count, Chain_Bond_Count, Rotatable_Bond_Count, Platt_Index, Randic_Index, Harary_Index, Hyper_Wiener_Index, Szeged_Index, Wiener_Index, Wiener_Polarity, H-Bond_Donor_Count, H-Bond_Donor_Sites, H-Bond_Acceptor_Count, H-Bond_Acceptor_Sites, Refractivity	74.8	0.401
OneRAttributeEval	Ranker	Hyper_Wiener_Index, H-Bond_Donor_Sites, Randic_Index, Harary_Index, Wiener_Index, Platt_Index, Aliphatic_Bond_Count, Szeged_Index, Wiener_Polarity, H-Bond_Donor_Count, Chain_Bond_Count, Aliphatic_Atom_Count, Chain_Atom_Count, H-Bond_Acceptor_Sites, pI, logD, logP, Bond_Count, Refractivity, Molecular_Polarizability, Atom_Count, Exact_Mass, Rotatable_Bond_Count, Molecular_Weight, H-Bond_Acceptor_Count, Aromatic_Atom_Count, Fused_Aromatic_Ring_Count, Largest_Ring_System_Size, Aromatic_Bond_Count, Ring_System_Count, Ring_Count, Smallest_Ring_Size, Largest_Ring_Size, Fragment_Count, Cyclomatic_Number, Smallest_Ring_System_Size, Asymetric_Atom_Count, Heteroaromatic_Ring_Count, Fused_Ring_Count, Aliphatic_Ring_Count, Carbo_Ring_Count, Carboaliphatic_Ring_Count, Carboaromatic_Ring_Count, Aromatic_Ring_Count, Stereo_Double_Bond_Count, Heteroaliphatic_Ring_Count, Hetero_Ring_Count, Chiral_Center_Count, Ring_Atom_Count, Ring_Bond_Count, Fused_Aliphatic_Ring_Count	74.8	0.401

Table A5: Continued

Attribute Evaluator	Search Method	Selected Attributes	J48 Accuracy	J48 RMS Error
PrincipalComponents	Ranker	Refractivity, Molecular_Polarizability, Bond_Count, Atom_Count, Wiener_Polarity, Randic_Index, Aliphatic_Bond_Count, Aliphatic_Atom_Count, Chain_Atom_Count, Chain_Bond_Count, Molecular_Weight, Exact_Mass, Platt_Index, H-Bond_Donor_Count, Rotatable_Bond_Count, Harary_Index, H-Bond_Acceptor_Count, H-Bond_Donor_Sites, H-Bond_Acceptor_Sites, logD, Szeged_Index, Wiener_Index, logP, Hyper_Wiener_Index, pI	74.8	0.401
ReliefFAttributeEval	Ranker	pI, Hyper_Wiener_Index, logP, H-Bond_Donor_Sites, logD, H-Bond_Donor_Count, H-Bond_Acceptor_Sites, Rotatable_Bond_Count, Bond_Count, Atom_Count, Molecular_Polarizability, Refractivity, Wiener_Polarity, Randic_Index, Aliphatic_Bond_Count, Chain_Bond_Count, Chain_Atom_Count, Aliphatic_Atom_Count, Molecular_Weight, Exact_Mass, Platt_Index, Harary_Index, H-Bond_Acceptor_Count, Szeged_Index, Wiener_Index	74.8	0.401
SVMAttributeEval	Ranker	Hyper_Wiener_Index, H-Bond_Donor_Sites, Randic_Index, Harary_Index, Wiener_Index, Platt_Index, Aliphatic_Bond_Count, Szeged_Index, Wiener_Polarity, H-Bond_Donor_Count, Chain_Bond_Count, Aliphatic_Atom_Count, Chain_Atom_Count, H-Bond_Acceptor_Sites, pI, logD, logP, Bond_Count, Refractivity, Molecular_Polarizability, Atom_Count, Exact_Mass, Rotatable_Bond_Count, Molecular_Weight, H-Bond_Acceptor_Count, Aromatic_Atom_Count, Fused_Aromatic_Ring_Count, Largest_Ring_System_Size, Aromatic_Bond_Count, Ring_System_Count, Ring_Count, Smallest_Ring_Size, Largest_Ring_Size, Fragment_Count, Cyclomatic_Number, Smallest_Ring_System_Size, Asymetric_Atom_Count, Heteroaromatic_Ring_Count, Fused_Ring_Count, Aliphatic_Ring_Count, Carbo_Ring_Count, Carboaliphatic_Ring_Count, Carboaromatic_Ring_Count, Aromatic_Ring_Count, Stereo_Double_Bond_Count, Heteroaliphatic_Ring_Count, Hetero_Ring_Count, Chiral_Center_Count, Ring_Atom_Count, Ring_Bond_Count, Fused_Aliphatic_Ring_Count	74.8	0.401

Table A5: Continued

Attribute Evaluator	Search Method	Selected Attributes	J48 Accuracy	J48 RMS Error
SymmetricalUncert AttibuteEval	Ranker	pI, logP, Harary_Index, Molecular_Polarizability, Bond_Count, Refractivity, Rotatable_Bond_Count, Atom_Count, H-Bond_Acceptor_Sites, Hyper_Wiener_Index, logD, Aliphatic_Bond_Count, Chain_Bond_Count, Chain_Atom_Count, Exact_Mass, Aliphatic_Atom_Count, Molecular_Weight, Wiener_Index, Randic_Index, Platt_Index, Szeged_Index, H-Bond_Donor_Count, Wiener_Polarity, H-Bond_Donor_Sites, H-Bond_Acceptor_Count	74.8	0.401
WrapperSubsetEval	GreedyStepwise	NO ATTRIBUTES SELECTED		

Table A6: Schema describing different properties of the various generations of PAMAM dendrimers[98, 142, 160-165].

Generation	Generation Classification	Surface Group	No. Surface Group	Molecular Weight (Da)	Diameter (nm)
G0	Full	$-NH_2$	4	517	1.5
G1	Full	$-NH_2$	8	1430	2.2
G1.5	Half	-COOH	16	2935	2.2
G2 ($-NH_2$)	Full	$-NH_2$	16	3256	2.9
G2 (-OH)	Full	-OH	16	3272	2.9
G2.5	Half	-COOH	32	6267	3.56
G3	Full	$-NH_2$	32	6909	3.6
G3.5	Half	-COOH	64	12931	3.8
G4	Full	$-NH_2$	64	14215	4.5
G4.5	Half	-COOH	128	26258	4.7

Table A7: Table listing all of the acronyms/abbreviations and their unabbreviated forms.

Acronym/Abbreviation	**Unabbreviated Form**
NLP	Natural language processing
PAMAM	Poly(amido amine)
ChiSquaredAttributeEval	Chi squared attribute evaluation
pI	Isoelectric point
LWL	Locally weighted learning
NBTree	Naïve Bayes tree
SMO	Sequential minimal optimization
DTNB	Decision table/Naïve Bayes

REFERENCES

1 de la Iglesia D, Maojo V, Chiesa S, *et al.* International efforts in nanoinformatics research applied to nanomedicine. *Methods Inf Med*. 2011;50(1):84-95.

2 Staggers N, McCasky T, Brazelton N, *et al.* Nanotechnology: the coming revolution and its implications for consumers, clinicians, and informatics. *Nurs Outlook*. 2008 Sep-Oct;56(5):268-74.

3 Zhou J, Liu J, Cheng CJ, *et al.* Biodegradable poly(amine-co-ester) terpolymers for targeted gene delivery. *Nat Mat*. 2012 Jan;11(1):82-90.

4 Zhao P, Wang H, Yu M, *et al.* Paclitaxel loaded folic acid targeted nanoparticles of mixed lipid-shell and polymer-core: In vitro and in vivo evaluation. *Eur J Pharm Biopharm*. 2012 Jun;81(2):248-56.

5 Thomas DG, Pappu RV, Baker NA. NanoParticle Ontology for cancer nanotechnology research. *J Biomed Inform*. 2011 Feb;44(1):59-74.

6 Moss DM, Siccardi M. Optimizing nanomedicine pharmacokinetics using physiologically based pharmacokinetics modelling. *Br J Pharmacol*. 2014 Sep;171(17):3963-79.

7 Elsaesser A, Howard CV. Toxicology of nanoparticles. *Adv Drug Deliv Rev*. 2012 Feb;64(2):129-37.

8 Fadeel B, Garcia-Bennett AE. Better safe than sorry: Understanding the toxicological properties of inorganic nanoparticles manufactured for biomedical applications. *Adv Drug Deliv Rev*. 2010 Mar 8;62(3):362-74.

9 Mukherjee SP, Davoren M, Byrne HJ. In vitro mammalian cytotoxicological study of PAMAM dendrimers - towards quantitative structure activity relationships. *Toxicol In Vitro*. 2010 Feb;24(1):169-77.

10 Adiseshaiah PP, Hall JB, McNeil SE. Nanomaterial standards for efficacy and toxicity assessment. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*. 2010;2(1):99-112.

11 Shahbazi MA, Santos HA. Improving oral absorption via drug-loaded nanocarriers: absorption mechanisms, intestinal models and rational fabrication. *Curr Drug Metab*. 2013 Jan;14(1):28-56.

12 Onoue S, Yamada S, Chan HK. Nanodrugs: pharmacokinetics and safety. *Int J Nanomedicine*. 2014;9:1025-37.

13 Hillaireau H, Couvreur P. Nanocarriers' entry into the cell: relevance to drug delivery. *Cell Mol Life Sci*. 2009 Sep;66(17):2873-96.

14 Smith PJ, Giroud M, Wiggins HL, et al. Cellular entry of nanoparticles via serum sensitive clathrin-mediated endocytosis, and plasma membrane permeabilization. *Int J Nanomedicine*. 2012;7:2045-55.

15 He B, Lin P, Jia Z, et al. The transport mechanisms of polymer nanoparticles in Caco-2 epithelial cells. *Biomaterials*. 2013 Aug;34(25):6082-98.

16 Jain K. *The Handbook of Nanomedicine*. 1st ed. Totowa, New Jersey: Humana; 2008.

17 Dawidczyk CM, Kim C, Park JH, et al. State-of-the-art in design rules for drug delivery platforms: lessons learned from FDA-approved nanomedicines. *J Control Release*. 2014 Aug 10;187:133-44.

18 U.S. National Science Foundation. Workshop on Nanoinformatics Strategies. 2007 [cited 2012; Available from: http://128.119.56.118/~nnn01/Workshop.html

19 NSTC/CoT/NSET. National Nanotechnology Initiative Strategic Plan (2014). 2014.

20 Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des*. 2007;13(34):3494-504.

21 de la Iglesia D, Cachau RE, Garcia-Remesal M, et al. Nanoinformatics knowledge infrastructures: bringing efficient information management to nanomedical research. *Comput Sci Discov*. 2013 Nov 27;6(1):014011.

22 Maojo V, Fritts M, de la Iglesia D, et al. Nanoinformatics: a new area of research in nanomedicine. *Int J Nanomedicine*. 2012;7:3867-90.

23 Maojo V, Fritts M, Martin-Sanchez F, et al. Nanoinformatics: developing new computing applications for nanomedicine. *Comput Sci Eng*. 2012 Jun 1;94(6):521-39.

24 Maojo V, Martin-Sanchez F, Kulikowski C, et al. Nanoinformatics and DNA-based computing: catalyzing nanomedicine. *Pediatric research*. 2010 May;67(5):481-9.

25 Thomas DG, Klaessig F, Harper SL, et al. Informatics and standards for nanomedicine technology. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*. 2011 Jun 30.

26 Panneerselvam S, Choi S. Nanoinformatics: emerging databases and available tools. *Int J Mol Sci*. 2014;15(5):7158-82.

27 Hunter L, Lu Z, Firby J, et al. OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*. 2008;9:78.

28 Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. *J Biomed Informatics*. 2011;44(1):163-79.

29 Chaussabel D. Biomedical literature mining: Challenges and solutions in the 'omics' era. *Amer J PharmacoGenomics*. 2004;4(6):383-93.

30 Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*. 2010;11(10):1467-89.

31 Lewinski NA, McInnes BT. Using natural language processing techniques to inform research on nanotechnology. *Beilstein J Nanotechnol*. 2015;6:1439-49.

32 Puzyn T, Leszczynska D, Leszczynski J. Toward the development of "nano-QSARs": advances and challenges. *Small*. 2009 Nov;5(22):2494-509.

33 Fourches D, Pu D, Tassa C, et al. Quantitative nanostructure-activity relationship modeling. *ACS Nano*. 2010 Oct 26;4(10):5703-12.

34 Shinoda W, DeVane R, Klein ML. Computer simulation studies of self-assembling macromolecules. *Curr Opin Struct Biol*. 2012 Apr;22(2):175-86.

35 Xiang TX, Anderson BD. Liposomal drug transport: a molecular perspective from molecular dynamics simulations in lipid bilayers. *Adv Drug Deliv Rev*. 2006 Nov 30;58(12-13):1357-78.

36 Johnson RR, Rego BJ, Johnson AT, et al. Computational study of a nanobiosensor: a single-walled carbon nanotube functionalized with the coxsackie-adenovirus receptor. *J Phys Chem B*. 2009 Aug 27;113(34):11589-93.

37 De Leo F, Sgrignani J, Bonifazi D, et al. Structural and dynamic properties of monoclonal antibodies immobilized on CNTs: a computational study. *Chemistry*. 2013 Sep 9;19(37):12281-93.

38 Stueker O, Ortega VA, Goss GG, et al. Understanding interactions of functionalized nanoparticles with proteins: a case study on lactate dehydrogenase. *Small*. 2014 May 28;10(10):2006-21.

39 Vacha R, Martinez-Veracoechea FJ, Frenkel D. Receptor-mediated endocytosis of nanoparticles of various shapes. *Nano Lett*. 2011 Dec 14;11(12):5391-5.

40 Li R, Chen R, Chen P, et al. Computational and experimental characterizations of silver nanoparticle-apolipoprotein biocorona. *J Phys Chem B*. 2013 Oct 31;117(43):13451-6.

41 du Toit LC, Pillay V, Choonara YE, et al. Patenting of nanopharmaceuticals in drug delivery: no small issue. *Recent Pat Drug Deliv Formul*. 2007;1(2):131-42.

42 Bielinska A, Kukowska-Latallo JF, Johnson J, et al. Regulation of in vitro gene expression using antisense oligonucleotides or antisense expression plasmids transfected using starburst PAMAM dendrimers. *Nucleic Acids Res*. 1996 Jun 1;24(11):2176-82.

43 Meltzer AD, Tirrell DA, Jones AA, et al. Chain Dynamics in Poly(amido amine) Dendrimers. A Study of 13C NMR Relaxation Parameters. *Macromolecules*. 1992;25:4541-8.

44 Kolhe P, Misra E, Kannan RM, et al. Drug complexation, in vitro release and cellular entry of dendrimers and hyperbranched polymers. *Int J Pharm*. 2003 Jun 18;259(1-2):143-60.

45 Wood KC, Little SR, Langer R, et al. A family of hierarchically self-assembling linear-dendritic hybrid polymers for highly efficient targeted gene delivery. *Angew Chem Int Ed Engl*. 2005 Oct 21;44(41):6704-8.

46 Greish K, Thiagarajan G, Herd H, et al. Size and surface charge significantly influence the toxicity of silica and dendritic nanoparticles. *Nanotoxicology*. 2012 Nov;6(7):713-23.

47 Thiagarajan G, Greish K, Ghandehari H. Charge affects the oral toxicity of poly(amidoamine) dendrimers. *Eur J Pharm Biopharm*. 2013 Jun;84(2):330-4.

48 Xu Q, Wang CH, Pack DW. Polymeric carriers for gene delivery: chitosan and poly(amidoamine) dendrimers. *Curr Pharm Des*. 2010 Jul;16(21):2350-68.

49 Yellepeddi VK, Kumar A, Palakurthi S. Surface modified poly(amido)amine dendrimers as diverse nanomolecules for biomedical applications. *Expert Opin Drug Deliv*. 2009;6(8):835-50.

50 Liu X, Tang K, Harper S, et al. Predictive modeling of nanomaterial exposure effects in biological systems. *Int J Nanomedicine*. 2013;8 Suppl 1:31-43.

51 Jensen LB, Mortensen K, Pavan GM, et al. Molecular characterization of the interaction between siRNA and PAMAM G7 dendrimers by SAXS, ITC, and molecular dynamics simulations. *Biomacromolecules*. 2010 Dec 13;11(12):3571-7.

52 Vasumathi V, Maiti PK. Complexation of siRNA with Dendrimer: A Molecular Modeling Approach. *Macromolecules*. 2010;43:8264-74.

53 Karatasos K, Posocco P, Laurini E, et al. Poly(amidoamine)-based dendrimer/siRNA complexation studied by computer simulations: effects of pH and generation on dendrimer structure and siRNA binding. *Macromol Biosci*. 2012 Feb;12(2):225-40.

54 Cao J, Zhang H, Wang Y, et al. Investigation on the interaction behavior between curcumin and PAMAM dendrimer by spectral and docking studies. *Spectrochim Acta A Mol Biomol Spectrosc*. 2013 May;108:251-5.

55 Castriciano MA, Romeo A, Angelini N, et al. Spectroscopic investigation and molecular modeling on porphyrin/PAMAM supramolecular adduct. *Photochem Photobiol*. 2011 Mar-Apr;87(2):292-301.

56 Avila-Salas F, Sandoval C, Caballero J, et al. Study of interaction energies between the PAMAM dendrimer and nonsteroidal anti-inflammatory drug using a distributed computational strategy and experimental analysis by ESI-MS/MS. *J Phys Chem B*. 2012 Feb 23;116(7):2031-9.

57 Ivanov AA, Jacobson KA. Molecular modeling of a PAMAM-CGS21680 dendrimer bound to an A2A adenosine receptor homodimer. *Bioorg Med Chem Lett*. 2008 Aug 1;18(15):4312-5.

58 Lee I, Majoros IJ, Williams CR, et al. Interactive Design Strategy for a Multi-Functional PAMAM Dendrimer-Based Nano-Therapeutic Using Computational Models and Experimental Analysis. *Journal Comput Theor Nanosci.* 2009;6(1):54-60.

59 Barata T, Teo I, Lalwani S, et al. Computational design principles for bioactive dendrimer based constructs as antagonists of the TLR4-MD-2-LPS complex. *Biomaterials.* 2011 Nov;32(33):8702-11.

60 Puzyn T, Rasulev B, Gajewicz A, et al. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat Nanotechnol.* 2011 Mar;6(3):175-8.

61 Jones AT, Gumbleton M, Duncan R. Understanding endocytic pathways and intracellular trafficking: a prerequisite for effective design of advanced drug delivery systems. *Adv Drug Deliv Rev.* 2003 11/14/;55(11):1353-7.

62 Gabizon A, Bradbury M, Prabhakar U, et al. Cancer nanomedicines: closing the translational gap. *The Lancet.* 2014 12/20/;384(9961):2175-6.

63 Weissleder R, Kelly K, Sun EY, et al. Cell-specific targeting of nanoparticles by multivalent attachment of small molecules. *Nat Biotechnol.* 2005 Nov;23(11):1418-23.

64 Kim SS, Rait A, Rubab F, et al. The clinical potential of targeted nanomedicine: delivering to cancer stem-like cells. *Mol Ther.* 2014 Feb;22(2):278-91.

65 Zheng W, Tropsha A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci.* 2000 Jan;40(1):185-94.

66 Shen M, Xiao Y, Golbraikh A, et al. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem.* 2003 Jul 3;46(14):3013-20.

67 Winkler DA, Burden FR, Yan B, et al. Modelling and predicting the biological effects of nanomaterials. *SAR QSAR Environ Res.* 2014;25(2):161-72.

68 Dragon Professional Software Package. 5.3 for Windows ed. Milano, Italy: Milano Chemometrics and QSAR Research Group; 2009.

69 Burden FR, Winkler DA. Optimum QSAR model feature selection using sparse Bayesian methods. *QSAR Comb Sci.* 2009;28:645-53.

70 Burden FR, Winkler DA. Robust QSAR models using Bayesian regularized neural networks. *J Med Chem.* 1999 Aug 12;42(16):3183-7.

71 Burden FR, Winkler DA. An optimal self-pruning neural network that performs nonlinear descriptor selection for QSAR. *QSAR Comb Sci.* 2009;28:1092-7.

72 Winkler DA, Burden FR. Robust QSAR models from novel descriptors and Bayesian regularized neural networks. *Mol Simul.* 2000;24:243-58.

73 Epa VC, Burden FR, Tassa C, et al. Modelling biological activities of nanoparticles. *Nano Lett*. 2012;12:5808-12.

74 Jones DE, Ghandehari H, Facelli JC. Predicting cytotoxicity of PAMAM dendrimers by molecular descriptors. *Beilstein J Nanotechnol*. 2015;6:1886-96.

75 Landsiedel R, Ma-Hock L, Kroll A, et al. Testing metal-oxide nanomaterials for human safety. *Adv Mater*. 2010 Jun 25;22(24):2601-27.

76 Sayes C, Ivanov I. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. *Risk Anal*. 2010 Nov;30(11):1723-34.

77 Devroye L, Gyorfi L, Lugosi G. *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag; 1996.

78 Stewart JJP. MOPAC2009. Stewart Computational Chemistry; 2009.

79 Liu R, Rallo R, George S, et al. Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small*. 2011 Apr 18;7(8):1118-26.

80 Horev-Azaria L, Baldi G, Beno D, et al. Predictive toxicology of cobalt ferrite nanoparticles: comparative in-vitro study of different cellular models using methods of knowledge discovery from data. *Part Fibre Toxicol*. 2013;10:32.

81 Quinlan R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers; 1993.

82 Toropova AP, Toropov AA, Rallo R, et al. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicol Environ Saf*. 2015 Feb;112:39-45.

83 Aha D, Kibler D, Albert MK. Instance-based learning algorithms. *Machine Learning*. 1991;6(1):37-66.

84 Breiman L. Bagging predictors. *Machine Learning*. 1996;24(2):123-40.

85 Quinlan RJ. Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence; 1992; Singapore; 1992. p. 343-8.

86 Cleary JG, Trigg LE. K*: An Instance-based Learner Using an Entropic Distance Measure. 12th International Conference on Machine Learning; 1995. p. 108-14.

87 Asharani PV, Lian Wu Y, Gong Z, et al. Toxicity of silver nanoparticles in zebrafish models. *Nanotechnology*. 2008 Jun 25;19(25):255102.

88 Harper SL, Carriere JL, Miller JM, et al. Systematic evaluation of nanomaterial toxicity: utility of standardized materials and rapid assays. *ACS Nano*. 2011 Jun 28;5(6):4688-97.

89 Witten I, Frank E, Hall M. *Data Mining: Practical Machine Learning Tools and Techniques*. 3 ed: Morgan Kaufmann Publishers; 2011.

90 Schoelkopf B, Burges C, Smola A. Advances in Kernel Methods - Support Vector Learning; 1998.

91 Frank E, Wang Y, Inglis S, et al. Using model trees for classification. *Machine Learning*. 1998;32(1):63-76.

92 Hall M, Frank E, Holmes G, et al. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. 2009;11(1):10-8.

93 Frank E, Hall M, Pfahringer B. Locally Weighted Naive Bayes. 19th Conference in Uncertainty in Artificial Intelligence; 2003. p. 249-56.

94 Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence; 1995.

95 Hall M, Frank E. Combining Naive Bayes and Decision Tables. 21st Florida Artificial Intelligence Society Conference (FLAIRS); 2008.

96 Kohavi R. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. Second International Conference on Knowledge Discovery and Data Mining; 1996. p. 202-7.

97 ChemAxon. Marvin 5.12.4. 2013.

98 El-Sayed M, Ginski M, Rhodes CA, et al. Influence of Surface Chemistry of Poly(Amidoamine) Dendrimers on Caco-2 Cell Monolayers. *J Bioact Compat Poly*. 2003;18(1):7-22.

99 Shalaby KS, Soliman ME, Casettari L, et al. Determination of factors controlling the particle size and entrapment efficiency of noscapine in PEG/PLA nanoparticles using artificial neural networks. *Int J Nanomedicine*. 2014;9:4953-64.

100 Leo E, Brina B, Forni F, et al. In vitro evaluation of PLA nanoparticles containing a lipophilic drug in water-soluble or insoluble form. *Int J Pharm*. 2004 Jun 18;278(1):133-41.

101 Husseini GA, Mjalli FS, Pitt WG, et al. Using artificial neural networks and model predictive control to optimize acoustically assisted Doxorubicin release from polymeric micelles. *Technol Cancer Res Treat*. 2009 Dec;8(6):479-88.

102 Makadia HK, Siegel SJ. Poly Lactic-co-Glycolic Acid (PLGA) as Biodegradable Controlled Drug Delivery Carrier. *Polymers*. 2011 Sep 1;3(3):1377-97.

103 Danhier F, Ansorena E, Silva JM, et al. PLGA-based nanoparticles: an overview of biomedical applications. *J Control Release*. 2012 Jul 20;161(2):505-22.

104 Szlek J, Paclawski A, Lau R, et al. Heuristic modeling of macromolecule release from PLGA microspheres. *Int J Nanomedicine*. 2013;8:4601-11.

105 Marvin cxcalc plugin, UK. 5.11 ed. Budapest, Hungary: ChemAxon.

106 Matsumura Y, Oda T, Maeda H. [General mechanism of intratumor accumulation of macromolecules: advantage of macromolecular therapeutics]. *Gan To Kagaku Ryoho*. 1987 Mar;14(3 Pt 2):821-9.

107 Jain RK. Barriers to drug delivery in solid tumors. *Sci Am*. 1994 Jul;271(1):58-65.

108 Boso DP, Lee SY, Ferrari M, et al. Optimizing particle size for targeting diseased microvasculature: from experiments to artificial neural networks. *Int J Nanomedicine*. 2011;6:1517-26.

109 Harashima H, Sakata K, Funato K, et al. Enhanced hepatic uptake of liposomes through complement activation depending on the size of liposomes. *Pharm Res*. 1994 Mar;11(3):402-6.

110 Ren J, Hong H, Song J, et al. Particle size and distribution of biodegradable poly-d,llactide- co-poly(ethylene glycol) block polymer nanoparticles prepared by nanoprecipitation. *J Appl Polym Sci*. 2005;98:1884-90.

111 Asadi H, Rostamizadeh K, Salari D, et al. Preparation of biodegradable nanoparticles of tri-block PLA-PEG-PLA copolymer and determination of factors controlling the particle size using artificial neural network. *J Microencapsul*. 2011;28(5):406-16.

112 Mukherjee B, Santra K, Pattnaik G, et al. Preparation, characterization and in-vitro evaluation of sustained release protein-loaded nanoparticles based on biodegradable polymers. *Int J Nanomedicine*. 2008;3(4):487-96.

113 Sattler KD. *Handbook of Nanophysics: Nanoparticles and Quantum Dots*. CRC Press; 2010.

114 Sinko PJ. *Martin's Physical Pharmacy and Pharmaceutical Sciences: Physical Chemical and Biopharmaceutical Principles in the Pharmaceutical Sciences*. Lippincott Williams & Wilkins; 2006.

115 Belbella AA, Vauthier C, Fessi H, et al. In vitro degradation of nanospheres from poly(D,L-lactides) of different molecular weights and polydispersities. *Int J Pharm*. 1996;129:95-102.

116 Schärtl W. *Light Scattering from Polymer Solutions and Nanoparticle Dispersions*. Berlin Heidelberg. Springer-Verlag; 2007.

117 Esmaeilzadeh-Gharehdaghi E, Faramarzi MA, Amini MA, et al. Processing/formulation parameters determining dispersity of chitosan particles: an ANNs study. *J Microencapsul*. 2014;31(1):77-85.

118 Dougherty ER, Hua J, Bittner ML. Validation of computational methods in genomics. *Current Genomics*. 2007;8(1):1-19.

119 Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. *Inf Proc Man*. 2006;42(1):155-65.

120 Galar M, Fernandez A, Barrenechea E, et al. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics*. 2012;42(4):463-84.

121 Tantra R, Oksel C, Puzyn T, et al. Nano(Q)SAR: Challenges, pitfalls and perspectives. *Nanotoxicology*. 2015;9(5):636-42.

122 Oksel C, Ma CY, Liu JL, et al. (Q)SAR modelling of nanomaterial toxicity: A critical review. *Particuology*. 2015;21:1-19.

123 Institute ANS. ANSI-NSP Launches Nanotechnology Standards Database. 2013 [cited 2013 09/30/2013]; Available from: http://www.ansi.org/news_publications/news_story.aspx?menuid=7&articleid=3681

124 nanoHUB.org. nanoHUB.org Online Simulation and More for Nanotechnology About Us. 2013 [cited 2013 09/30/2013]; Available from: http://nanohub.org/about

125 National Cancer Institute. caNanoLab. 2011 [cited 2011; Welcome to the cancer Nanotechnology Laboratory (caNanoLab) portal. caNanoLab is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical and in vitro assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.]. Available from: https://cananolab.nci.nih.gov/caNanoLab/

126 Garcia-Remesal M, Garcia-Ruiz A, Perez-Rey D, et al. Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature. *Biomed Res Int*. 2013;2013:410294.

127 Cunningham H, al. e. Text Processing with GATE. University of Sheffield Department of Computer Science; 2011.

128 Zaremba S, Ramos-Santacruz M, Hampton T, et al. Text-mining of PubMed abstracts by natural language processing to create a public knowledge base on molecular mechanisms of bacterial enteropathogens. *BMC Bioinformatics*. 2009;10:177.

129 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: John Wiley; 1981.

130 Yang Y. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*. 1999;1(1-2):69-90.

131 Jones DE, Igo S, Hurdle J, et al. Automatic extraction of nanoparticle properties using natural language processing: NanoSifter an application to acquire PAMAM dendrimer properties. *PloS One*. 2014;9(1):e83932.

132 Kohavi R. The Power of Decision Tables. 8th European Conference on Machine Learning; 1995. p. 174-89.

133 Goldberg DS, Vijayalakshmi N, Swaan PW, et al. G3.5 PAMAM dendrimers enhance transepithelial transport of SN38 while minimizing gastrointestinal toxicity. *J Control Release*. 2011 Mar 30;150(3):318-25.

134 Jevprasesphant R, Penny J, Jalal R, et al. The influence of surface modification on the cytotoxicity of PAMAM dendrimers. *Int J Pharm*. 2003 Feb 18;252(1-2):263-6.

135 Ke W, Zhao Y, Huang R, et al. Enhanced Oral Bioavailability of Doxorubicin in a Dendrimer Drug Delivery System. *J Pharm Sci*. 2008;97(6):2208-16.

136 Kitchens KM, Foraker AB, Kolhatkar RB, et al. Endocytosis and interaction of poly (amidoamine) dendrimers with Caco-2 cells. *Pharm Res*. 2007 Nov;24(11):2138-45.

137 Kitchens KM, Kolhatkar RB, Swaan PW, et al. Endocytosis inhibitors prevent poly(amidoamine) dendrimer internalization and permeability across Caco-2 cells. *Mol Pharm*. 2008 Mar-Apr;5(2):364-9.

138 Kolhatkar RB, Kitchens KM, Swaan PW, et al. Surface acetylation of polyamidoamine (PAMAM) dendrimers decreases cytotoxicity while maintaining membrane permeability. *Bioconjug Chem*. 2007 Nov-Dec;18(6):2054-60.

139 Najlah M, Freeman S, Attwood D, et al. In vitro evaluation of dendrimer prodrugs for oral drug delivery. *Int J Pharm*. 2007 May 4;336(1):183-90.

140 Pisal DS, Yellepeddi VK, Kumar A, et al. Permeability of surface-modified polyamidoamine (PAMAM) dendrimers across Caco-2 cell monolayers. *Int J Pharm*. 2008 Feb 28;350(1-2):113-21.

141 Schilrreff P, Mundina-Weilenmann C, Romero EL, et al. Selective cytotoxicity of PAMAM G5 core--PAMAM G2.5 shell tecto-dendrimers on melanoma cells. *Int J Nanomedicine*. 2012;7:4121-33.

142 Sweet DM, Kolhatkar RB, Ray A, et al. Transepithelial transport of PEGylated anionic poly(amidoamine) dendrimers: implications for oral drug delivery. *J Control Release*. 2009 Aug 19;138(1):78-85.

143 Teow HM, Zhou Z, Najlah M, et al. Delivery of paclitaxel across cellular barriers using a dendrimer-based nanocarrier. *Int J Pharm*. 2013 Jan 30;441(1-2):701-11.

144 Najlah M, Freeman S, Attwood D, et al. Synthesis and Assessment of First-Generation Polyamidoamine Dendrimer Prodrugs to Enhance the Cellular Permeability of P-gp Substrates. *Bioconjug Chem*. 2007;18:937-46.

145 ChemAxon, Berry I, Ruyts B. 2012. Future-proofing Cheminformatics Platforms. Available from: https://www.chemaxon.com/publication/future-proofing-cheminformatics-platforms/

146 Maimon O, Rokach L. *The Data Mining and Knowledge Discovery Handbook*. New York: Springer-Verlag; 2005.

147 Wang B, Navath RS, Menjoge AR, et al. Inhibition of bacterial growth and intramniotic infection in a guinea pig model of chorioamnionitis using PAMAM dendrimers. *Int J Pharm*. 2010 Aug 16;395(1-2):298-308.

148 Noach AB, Kurosaki Y, Blom-Roosemalen MC, et al. Cell-polarity dependent effect of chelation on the paracellular permeability of confluent caco-2 cell monolayers. *Int J Pharm*. 1993;90(3):229-37.

149 Amsterdam A, Jamieson JD. Studies on dispersed pancreatic exocrine cells. I. Dissociation technique and morphologic characteristics of separated cells. *Journal Cell Biol*. 1974 Dec;63(3):1037-56.

150 Hubbard D, Ghandehari H, Brayden DJ. Transepithelial transport of PAMAM dendrimers across isolated rat jejunal mucosae in ussing chambers. *Biomacromolecules*. 2014 Aug 11;15(8):2889-95.

151 Hubbard D, Enda M, Bond T, et al. Transepithelial Transport of PAMAM Dendrimers Across Isolated Human Intestinal Tissue. *Mol Pharm*. 2015 Nov 2;12(11):4099-107.

152 Case DA, Babin V, Berryman JT, et al. AMBER 14. University of California, San Francisco; 2014.

153 Jorgensen WL, Chandrasekhar J, Madura JD, et al. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79(2):926-35.

154 Tomita M, Hayashi M, Awazu S. Absorption-enhancing mechanism of EDTA, caprate, and decanoylcarnitine in Caco-2 cells. *J Pharm Sci*. 1996 Jun;85(6):608-11.

155 Vllasaliu D, Shubber S, Garnett M, et al. Evaluation of calcium depletion as a strategy for enhancement of mucosal absorption of macromolecules. *Biochem Biophys Res Commun*. 2012 Feb 3;418(1):128-33.

156 Roe DR, Cheatham TE, 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput*. 2013 Jul 9;9(7):3084-95.

157 Wieser ME. Atomic weights of the elements 2005 (IUPAC Technical Report)". *Pure Appl Chem*; 2006.

158 Audi G. The 1995 update to the atomic mass evaluation. *Nuclear Physics*. 1995;4:409-80.

159 Viswanadhan VN, Ghose AK, Revankar GR, et al. Atomic physiochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J Chem Inf Comput Sci*. 1989;29:163-72.

160 Tomalia DA, Naylor AM, Goddard WA. Starburst Dendrimers: Molecular-Level Control of Size, Shape, Surface Chemistry, Topology, and Flexibility from Atoms to Macroscopic Matter. *Angew Chem Int Ed Engl*. 1990;29(2):138-75.

161 Prieto MJ, del Rio Zabala NE, Marotta CH, et al. Optimization and in vivo toxicity evaluation of G4.5 PAMAM dendrimer-risperidone complexes. *PloS One*. 2014;9(2):e90393.

162 Vázquez-Olmos A, Díaz D, Rodríguez-Gattorno G, et al. Activation of CdS nanoparticles by metallic ions and their selective interactions with PAMAM dendrimers. *Colloid Polym Sci*. 2003;282(9):957-64.

163 Satoh K, Yoshimura T, Esumi K. Effects of Various Thiol Molecules Added on Morphology of Dendrimer-Gold Nanocomposites. *J Coll Interface Sci*. 2002;255:312-22.

164 Kitchens KM, Kolhatkar RB, Swaan PW, et al. Transport of poly(amidoamine) dendrimers across Caco-2 cell monolayers: Influence of size, charge and fluorescent labeling. *Pharm Res*. 2006 Dec;23(12):2818-26.

165 Devarakonda B, Hill RA, de Villiers MM. The effect of PAMAM dendrimer generation size and surface functional group on the aqueous solubility of nifedipine. *Int J Pharm*. 2004 Oct 13;284(1-2):133-40.